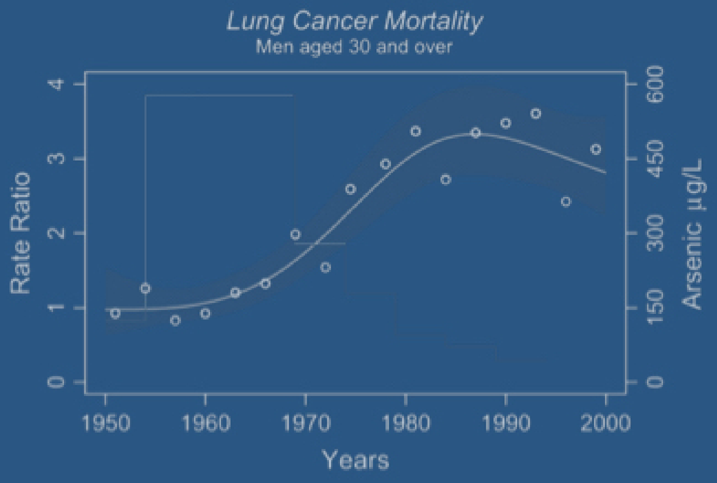
$$L_h(t) = \prod_{j=1}^{m} p_{i_{j-1},i_j}(\Delta t_j; z)\{f_{i_m}(\Delta t_{m+1}\,|\,z)\}^{\delta}\{S_{i_m}(\Delta t_{m+1}\,|\,z)\}^{1-\delta}$$

Contributions to Biostatistics

and Clinical & Epidemiological Studies



Guillermo Marshall, PhD

$$D^{(\nu+1)} = \frac{1}{N}\sum_{i=1}^{N}\left(\tilde{\beta}_i\tilde{\beta}_i' + D^{(\nu)} - D^{(\nu)}\tilde{Z}_i' W_i^{(\nu)}\tilde{Z}_i D^{(\nu)}\right)$$

Table of Contents

Preface

Chapter 1

Multi–state Markov Regression Models and Disease Progression

Comments by Wensheng Guo, Ph.D, FASA, FIMS [1]

I would like to congratulate Dr. Guillermo Marshall for the wonderful book and his fundamental contributions to multistate Markov transition models. I would also like to take this opportunity to express my deepest gratitude towards Guillermo for introducing me to this field and providing me with the opportunity to work with him on discrete time Markov transition models when I was just a first-year graduate student at University of Colorado. This work inspired me to pursue an academic career and to further develop methodology in state space models, which are highly related to Markov transition models.

The first part of this chapter was based on Guillermo's dissertation work, in which he extended the multistate Markov transition model of Kay (1986) to 1) to account for when the time–to-event is exactly observed, 2) to include covariates effects through various regression models, and 3) to approximate a continuous time non-homogeneous Markov model with a piecewise homogeneous Markov model. This led to a new framework of flexible models that are easy to interpret and can make direct inference on covariates effects on time-to-event. While only a small part of dissertation was published in the 1995 Statistics Medicine paper, the impact of that paper alone is already enormous. That paper has been cited 208 times, based on Goggle Scholar. It has been widely applied in many fields such as various disciplines of medicines, epidemiology, genetics, economics, and management. Because of the popularity of the multistate Markov transition modes, there have been many related user-friendly software packages developed in R or Matlab in recent years.

The second part of the chapter describes the unpublished work on discrete time approximation to continuous time Markov transition model, in which I had the opportunity to participate. While the piecewise homogeneous approximation to the continuous time non-homogeneous Markov transition model leads to a computable solution, the computational demand is heavy due to the need to convert the corresponding transition intensity matrix to the transitional probability matrix for each observed transition (See chapter 1.1 for details). Guillermo's clever idea was to replace the intensity function by a "1-step" transition probability, and the transition probability of a given interval can be easily calculated from the "1-step" transition probability matrix. As long as the "step" is small enough, the results can be almost as accurate as the continuous time Model. This contrasts with some of discrete time Markov models published later that directly model the transition probability of an observed transition. If the observed data are unequally spaced, directly modeling the transition probability would require very strong assumptions. A full Markov transition model usually requires a large number of parameters that can grow quickly with the number of covariates. To reduce the number of parameters when the outcome is ordinal, we impose a testable proportional odds assumption, resulting in substantial parsimony in parameterization. This paper was not published because Guillermo went back to Chile, and I transferred to another school. Fortunately, the software package was published in time and the proposed method has been applied in various fields.

Three decades after Guillermo finished his dissertation, multistate Markov transition model is still an active research area. This work has been extended in various in directions, such as classification and clustering, Semi-Markov and hidden Markov setting, joint modeling of discrete longitudinal data and time-to-event, and semiparametric Markov transition model, just to name a few. Bayesian inference and MCMC procedures were also developed for computation when maximum likelihood estimate is difficult to obtain. With expansion of the computational power,

[1] Wensheng Guo is Professor of Biostatistics at University of Pennsylvania

I expect this area will grow even faster. And I also envision some of the recent machine learning techniques can be introduced to enable various nonlinear and nonparametric extensions.

Article 1.1

Multi-State Markov Models in Survival Analysis

Guillermo Marshall

University of Colorado Health Sciences Center, Denver, U.S.A.

Abstract. This work discusses a general k-state Markov model when the process is observed at irregular intervals and the exact transition times are not available. This represents an extension of previous work by (13). The Markov model is a generalization of parametric models in survival analysis and direct relations between the transition probabilities and transition intensities with survival functions are derived. Of particular interest is the application of this model to survival studies. An exact likelihood function is proposed to replace Kay's approximation when the exact transition time to the absorbing state is observed. A model with covariables and its likelihood function is provided in detail. Natural extensions to non-homogenous Markov models are discussed, and a model for time dependent covariables is proposed. As an example, a Markov model is applied to longitudinal data from a study of young patients with diabetes in order to describe the natural course of diabetic retinopathy.

Keywords: Multi-state models; Markov processes; Survival analysis; Longitudinal data; Time-dependent covariables. [1]

1 Introduction

In recent years, Markov and semi-Markov models have become important tools to describe and help understand the progression and regression of multi-state diseases such as cancer, HIV infection, leukemia, diabetes, and many others. These models have been used by many authors in the area of biomedical sciences to find possible markers for the transition from stable states to the accelerated phase and/or the irreversible (absorbing) state of a disease, and also to describe the natural course of these diseases. Klein, Klotz and Grever (14) use a three-state Semi-Markov Model in a study of patients with chronic myelogenous leukemia to analyze the effect of elevated blood levels of adenosine deaminase as a marker for transition from stable disease to blast crisis, and then to death. Kay (13) proposed a methodology to fit a general k disease state Markov model in continuous time with application to the analysis of cancer markers in survival studies. Longini et al. (15) use a sub-model to describe the distribution of the incubation period for AIDS patients.

Important contributions in the area of continuous-time Markov models have been due to work in areas such as social sciences, demography and engineering. Schoen and Kenneth (1979) use Markov models to estimate increment-decrement life tables with applications to marital-status.

[1] This article was written based on the main results of my Doctoral Dissertation under the supervision of Richard H. Jones from University of Colorado, 1990. Although this article was never published in the present form, some of the most important results were published as an entry in the **Encyclopedia of Biostatistics** with the title Predictive Modelling for Prognosis, Article 1.4 of this book and in Article 1.5

Kalbfleisch and Lawless (1985) introduce a continuous-time Markov model to analyze panel data, and Kalbfleisch, Lawless and Vollmer (1983) propose methods to estimate the parameters of this model from aggregate data. Madsen, Spliid and Thyregod (1985) applied discrete and continuous-time Markov models to describe the variation of cloud cover at an airport.

This paper discusses a general k-state Markov model in which the exact transition times are not observed, and represents an extension of the work of Kay (13). Of particular relevance is the extension of the relation between continuous-time Markov models and survival analysis functions. Multi-state Markov models represent a generalization of parametric models in survival analysis to the analysis of data concerning multiple events.

Multi-state Markov models are extended to include covariables in the transition intensities as in proportional hazard models (3) , and a general model based on a non-homogeneous Markov process is explored. Models for time- dependent covariables are also proposed. An important application of these models is discussed and analyzed. Data from a longitudinal study in young patients with diabetes from the Barbara Davis Center for Childhood Diabetes, University of Colorado Health Sciences Center, are used to determine factors affecting the natural course of diabetic retinopathy.

This paper is organized in four sections. Section 2 presents a brief review of the most relevant aspects of Markov processes in continuous time related to the methodology of the multi-state Markov models. Section 3 introduces the multi-state Markov model with covariables for a partially observed time-homogeneous Markov process. An exact likelihood function is proposed to replace Kay's approximation when the transition to the absorbing state is observed, and a complete relation between the multi-state model and survival analysis is derived. Finally, in Section 4, a non-homogeneous model is introduced.

2 Continuous-time Markov Processes

Introduction

Suppose we observe a continuous-time stochastic process $\{X(t), t \geq 0\}$ with a finite number of values in $E = \{1, 2, \ldots, k\}$ called states. We say that $\{X(t)\}$ is a continuous-time Markov process if for all times $t > s > u > 0$, and for any states i, j and $h \in E$,

$$pr\{X(t) = j | X(s) = i, X(u) = h\} = pr\{X(t) = j | X(s) = i\}. \qquad (1)$$

This conditional probability represents the probability of a transition from the state i at time s to the state j at time t, and it is denotated as $p_{ij}(s,t)$. These transition probabilities have the basic properties, $0 \leq p_{ij}(s,t) \leq 1$, $p_{ii}(t,t) = 1$, $p_{ij}(t,t) = 0$ if $j \neq i$ and

$$\sum_{j=1}^{k} p_{ij}(s,t) = 1.$$

For any time τ in the interval (s,t) the transition probability $p_{ij}(s,t)$ can be written using the Chapman- Kolmogorov equation as

$$p_{ij}(s,t) = \sum_{v=1}^{k} p_{iv}(s,\tau) p_{vj}(\tau,t) \ .$$

This equation can be written in matrix notation as $\mathbf{P}(s,t) = \mathbf{P}(s,\tau)\mathbf{P}(\tau,t)$, where $\mathbf{P}(s,t)$ is the transition probability matrix of dimension $k \times k$ with elements $p_{ij}(s,t)$.

The Markov process $X(t)$ can also be characterized in terms of the transition intensities,

$$q_{ij}(t) = \lim_{dt \to 0} \frac{Pr\{\ X(t+dt) = j \mid X(t) = i\ \}}{dt}\ , \qquad i \neq j$$

which represent instantaneous transition rates between the different states. For mathematical convenience, we define

$$q_{ii}(t) = -\sum_{j \neq i}^{k} q_{ij}(t).$$

The transition probability $p_{ij}(s,t)$ satisfies the Kolmogorov forward differential equations

$$\frac{dp_{ij}(s,t)}{dt} = \sum_{v=1}^{k} p_{iv}(s,t) q_{vj}(t),$$

or in matrix notation

$$\frac{d\mathbf{P}(s,t)}{dt} = \mathbf{P}(s,t)\mathbf{Q}(t) \tag{2}$$

with the initial condition $\mathbf{P}(t,t) = \mathbf{I}$, where $\mathbf{I}$ is the $k \times k$ identity matrix.

Time-homogeneous Markov processes

Important mathematical simplifications are obtained by assuming that the process $\{X(t), t \geq 0\}$ is homogeneous in time. The consequences of this assumption are that the transition intensities between the different states $q_{ij}(t)$ are constant over time, and the transition probabilities $p_{ij}(s,t)$ depend only on the time differences $t-s$. Equation (2) reduces to a system of differential equations with constant coefficients,

$$\frac{d\mathbf{P}(t-s)}{dt} = \mathbf{P}(t-s)\mathbf{Q},$$

for which the closed form solution is

$$\mathbf{P}(t-s) = e^{\mathbf{Q}(t-s)} = \sum_{n=0}^{\infty} \frac{\{\mathbf{Q}(t-s)\}^n}{n!}.$$

If $\mathbf{Q}$ has distinct eigenvalues, $\rho_1, \rho_2, \ldots, \rho_k$, and $\mathbf{A}$ is a square matrix where the jth column is the eigenvector associated with ρ_j, then we can calculate $\mathbf{P}(t-s)$ as

$$\mathbf{P}(t-s) \ = \ \mathbf{A}\mathrm{diag}\{\ e^{\rho_1(t-s)}, e^{\rho_2(t-s)}, \ldots, e^{\rho_k(t-s)}\ \}\mathbf{A}^{-1}\ . \tag{3}$$

Individual transition probabilities can be calculated, for any value of $t-s$, as

$$p_{ij}(t-s) = \sum_{v=1}^{k} a_{iv} e^{\rho_v(t-s)} a^{vj}, \tag{4}$$

where a_{ij} and a^{ij} represent the elements (i,j) of the matrices $\mathbf{A}$ and $\mathbf{A}^{-1}$. For more details about continuous-time Markov processes see Cox and Miller (2) and Chiang (1).

Non-homogeneous Markov processes

Non-homogeneous Markov processes are natural generalizations when transition rates change over time. Although this extension is important, the complexities of the mathematics and the computational difficulties are important obstacles in practical applications. Although a solution of the system of the differential equations (2) exists for a general form of Q(t), a closed form solution exists only for particular cases. If $\mathbf{Q}(t)$ is triangular, for example, processes with unidirectional transition like HIV infection, then a solution of (2) can be found by sequential integration. A good discussion of this method can be found in Raman and Chiang (1973) and Davies (1985).

An alternative method for finding a solution of (2) for a general $\mathbf{Q}(t)$ is to approximate the intensity matrix $\mathbf{Q}(t)$ by dividing the period of follow-up into K intervals $[\tau_1, \tau_2), [\tau_2, \tau_3), \ldots, [\tau_K, \infty)$ and assuming that the intensity matrix is constant during each interval. Hence, the intensity matrix $\mathbf{Q}(t)$ is approximated by a step function

$$\mathbf{Q}(t) = \mathbf{Q}_l, \quad t \in [\tau_l, \tau_{l+1}),$$

for $l = 1, 2, \ldots, K$. Local solutions of the Kolmogorov differential equations are

$$\mathbf{P}(\tau_l, t) = e^{\mathbf{Q}_l(t-\tau_l)},$$

where $t \in [\tau_l, \tau_{l+1})$, subject to the condition $\mathbf{P}(\tau_l, \tau_l) = \mathbf{I}$, for $l = 1, 2, \ldots, K$.

Using the Chapman-Kolmogorov equation and the above local solution, we can find a global solution as follows. For any two times s and t with $s \leq t$

$$\begin{aligned}\mathbf{P}(s,t) &= \mathbf{P}(s, \tau_{l+1})\mathbf{P}(\tau_{l+1}, \tau_{l+2}) \cdots \mathbf{P}(\tau_m, t) \\ &= e^{\mathbf{Q}_l(\tau_{l+1}-s)} e^{\mathbf{Q}_{l+1}(\tau_{l+2}-\tau_{l+1})} \cdots e^{\mathbf{Q}_m(t-\tau_m)},\end{aligned} \tag{5}$$

where $\tau_l < s \leq \tau_{l+1}$, $\tau_m < t \leq \tau_{m+1}$.

In the special case when the transition intensity matrices $\mathbf{Q}_1, \ldots, \mathbf{Q}_K$ permute, $\mathbf{Q}_i\mathbf{Q}_j = \mathbf{Q}_j\mathbf{Q}_i$, the expression (5) can be reduced to the exponent of the sum of the $\mathbf{Q}_i$ matrices. In general, if for any two times s and t, $\mathbf{Q}(s)\mathbf{Q}(t) = \mathbf{Q}(t)\mathbf{Q}(s)$, (2) has a closed form solution,

$$\mathbf{P}(s,t) = e^{\int_s^t \mathbf{Q}(u)du} \;. \tag{6}$$

A basic example of a transition intensity matrix that satisfies this is $\mathbf{Q}(t) = \mathbf{Q}h(t)$, where $h(t)$ is any positive function of t. In this case the closed form solution (6) reduces to

$$\mathbf{P}(s,t) = e^{\mathbf{Q}\int_s^t h(u)du} \;.$$

If h(t) = 1 we have the solution of the time-homogeneous Markov process discussed in the previous section. For more details about the system of equations (2) and methods to obtain its solutions see Gantmacher (7) and Hochstadt (9).

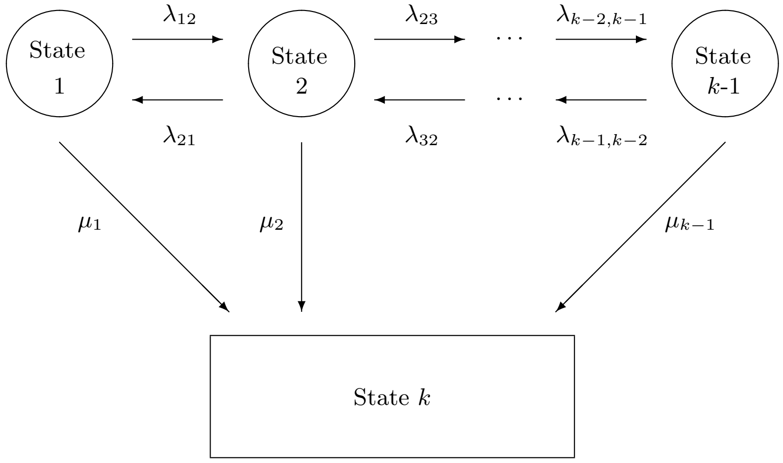
Computation of the transition probability matrix $\mathbf{P}(s,t)$ using equation (5) can be obtained using canonical decompositions for each of the terms $\exp\{\mathbf{Q}_l(\tau_{l+1} - \tau_l)\}$ as in (3) for the time-homogeneous case. The element (i,j) of the resulting product of matrices is $p_{ij}(s,t)$. The computing cost of calculating $p_{ij}(s,t)$ will depend on how small the intervals (τ_l, τ_{l+1}) are chosen, and the length of the intervals between observations of the process.

3 Time-Homogeneous Markov Models

The Basic Model

Suppose we have a model with $k-1$ transient disease states $j=1,\ldots,k-1$ and a single absorbing state $j=k$, see Figure 1. The transient states are assumed to be ordered according to j and instantaneous transitions can take place from state j to the adjoining states $j-1$ or $j+1$. Transitions can also take place from any of the transient states to the absorbing state k.

Fig. 1. A multi-state model with $k-1$ transient states and one absorbing state. The model has a total of $3k-5$ parameters, $2k-4$ λ's and $k-1$ μ's.



Submodels can be obtained by eliminating some of the parameters when some of the transitions are theoretically impossible or are unlikely to be observed during the study time. Longini et al. (15) use a submodel to describe the incubation period of AIDS with only progression transitions to the adjoining states.

For the model in Figure 1 the transition intensity matrix $\mathbf{Q}$ can be written as

$$\mathbf{Q}=\begin{pmatrix} -(\mu_1+\lambda_{12}) & \lambda_{12} & \cdots & 0 & \mu_1 \\ \lambda_{21} & -(\mu_2+\lambda_{21}+\lambda_{23}) & \cdots & 0 & \mu_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & -(\mu_{k-1}+\lambda_{k-1,k-2}) & \mu_{k-1} \\ 0 & 0 & \cdots & 0 & 0 \end{pmatrix}.$$

When equally spaced observations are available, and where a discrete time Markov model can be considered, a continuous- time Markov model will be more parsimonious. This model has $3k-5$ different parameters in contrast with a discrete time model with $k(k-1)$ independent parameters.

Given the form of the transition intensity matrix $\mathbf{Q}$, since all the λ's and μ's are non–negative, the eigenvalues of $\mathbf{Q}$ have negative real parts (Cox and Miller, 1965), except that $\rho_k = 0$. A necessary condition for all the eigenvalues of $\mathbf{Q}$ to be real is that

$$\lambda_{i,i+1}\lambda_{i+1,i} \geq 0 \quad , \quad i = 1, \ldots, k-1 \ .$$

The proof of this follows from the fact that the eigenvalue associated with the kth row of $\mathbf{Q}$ is always equal to zero, and that the $(k-1) \times (k-1)$ upper left sub- matrix of $\mathbf{Q}$ is a tri–diagonal matrix. If a tri- diagonal matrix satisfies the above restrictions then it is said to be symmetrizable and have a real eigensystem (19).

Let $\boldsymbol{R}$ be the $(k-1)\times(k-1)$ upper left sub-matrix of $\mathbf{Q}$. $\boldsymbol{R}$ is symmetrizable by a similarity transformation with diagonal matrix $\mathbf{D}$. If we define $\mathbf{D}$ as $d_{11} = 1$ and $d_{ii} = (\lambda_{21}\lambda_{32}\cdots\lambda_{i,i-1})/(\lambda_{12}\lambda_{23}\cdots\lambda_{i-1,i})$ for $i = 2, \ldots, k-1$, then

$$\mathbf{T} = \mathbf{D}^{-\frac{1}{2}}\boldsymbol{R}\mathbf{D}^{\frac{1}{2}}$$

is a tri-diagonal symmetric matrix with elements $t_{ii} = q_{ii}$ and $t_{i,i+1} = t_{i+1,i} = (\lambda_{i,i+1}\lambda_{i+1,i})^{\frac{1}{2}}$.

If a $\lambda_{i,i+1}$ or $\lambda_{i+1,i}$ is zero, the eigenvalues of $\boldsymbol{R}$ are the eigenvalues of a number of smaller tri-diagonal matrices, so that this case does not cause difficulties. Once we transform $\boldsymbol{R}$ to $\mathbf{T}$, we can use standard routines for finding the eigensystem in tri-diagonal symmetric matrices (10; 19)

Regression Models

An extended model can be developed by introducing covariables in the basic model discussed in the previous section. As in the proportional hazard model (Cox, 1972), a proportional transition intensity model can be used for each possible transition of the process.

Suppose that each individual under study has an associated vector of covariables $\boldsymbol{z}' = (z_1, z_2, z_3, ..., z_p)$, then for a given $\boldsymbol{z}$, we assume that the Markov process is homogeneous with an intensity matrix $\mathbf{Q}(\boldsymbol{z})$ with elements

$$\lambda_{ij}(\boldsymbol{z}) = \lambda_{ij}\psi(\boldsymbol{z};\boldsymbol{\beta}_{ij}) \ \text{ and } \ \mu_i(\boldsymbol{z}) = \mu_i\psi(\boldsymbol{z};\boldsymbol{\beta}_{ij}),$$

where λ_{ij} and μ_i represent the baseline transition rates, and $\boldsymbol{\beta}_{ij}$ represents the vector of regression coefficients associated with $\mathbf{z}$.

Different parametric forms of ψ can be considered such as the linear form $\psi(\boldsymbol{z};\boldsymbol{\beta}_{ij}) = 1 + \boldsymbol{\beta}'_{ij}\boldsymbol{z}$, the logistic, $\psi(\boldsymbol{z};\boldsymbol{\beta}_{ij}) = \log(1 + e^{\boldsymbol{\beta}'_{ij}\boldsymbol{z}})$, or the log linear form $\psi(\boldsymbol{z};\boldsymbol{\beta}_{ij}) = e^{\boldsymbol{\beta}'_{ij}\boldsymbol{z}}$. When referring to this model we will assume a log linear form since it is easy to apply and is always positive. For a further discussion of these parametric forms see Cox and Oakes (1984).

The solution of the Kolmogorov forward system of equations for this model is $\mathbf{P}(t-s;\boldsymbol{z}) = \exp\{\mathbf{Q}(\boldsymbol{z})(t-s)\}$, and a spectral decomposition of $\mathbf{Q}(\boldsymbol{z})$ can be used to calculate $\mathbf{P}(t-s;\boldsymbol{z})$ for any value of $t-s$ from equation (3). Individual transition probabilities can be evaluated for any value of $t-s$, using an expression similar to (4),

$$p_{ij}(t-s;\boldsymbol{z}) = \sum_{v=1}^{k} a_{iv}(\boldsymbol{z})e^{\rho_v(\boldsymbol{z})(t-s)}a^{vj}(\boldsymbol{z}), \tag{7}$$

where $a_{ij}(\boldsymbol{z})$ and $a^{ij}(\boldsymbol{z})$ represent the elements (i,j) of the matrices $\mathbf{A}(\boldsymbol{z})$ and $\mathbf{A}(\boldsymbol{z})^{-1}$, and where $\rho_1(\boldsymbol{z}), \rho_2(\boldsymbol{z}), \ldots, \rho_k(\boldsymbol{z})$ and $\mathbf{A}(\boldsymbol{z})$ are the eigenvalues and the matrix of eigenvectors of the transition intensity matrix $\mathbf{Q}(\boldsymbol{z})$.

Survival Analysis

A point of major interest in practical applications is the relationship between this Markov model and survival analysis functions, including the survival function, the hazard function, the median lifetime, the mean lifetime and the residual mean lifetime. Let T be a random variable which represents the lifetime of individuals in a homogeneous population. The survival function, given that the process is in state i at time $s = 0$, $S_i(t) = pr\{T > t | X(0) = i\}$, for a subject with covariables $\mathbf{z}$, is

$$S_i(t|\boldsymbol{z}) = 1 - p_{ik}(t;\boldsymbol{z}),$$

where $p_{ik}(t;\boldsymbol{z})$ is the element (i,k) of the transition probability matrix $\mathbf{P}(t;\boldsymbol{z})$. The density function, expressed in terms of the survival function, $f_i(t) = -dS_i(t)/dt$, for a subject with covariables $\mathbf{z}$, is

$$f_i(t|\boldsymbol{z}) = \sum_{j=1}^{k-1} p_{ij}(t;\boldsymbol{z})\mu_j(\boldsymbol{z}) \ .$$

The hazard function $h_i(t|\boldsymbol{z}) = f_i(t|\boldsymbol{z})/S_i(t|\boldsymbol{z})$ as a function of the transition probabilities and intensities is

$$h_i(t|\boldsymbol{z}) = \sum_{j=1}^{k-1} \frac{p_{ij}(t;\boldsymbol{z})}{\sum_{v=1}^{k-1} p_{iv}(t;\boldsymbol{z})}\mu_j(\boldsymbol{z}),$$

which represents a weighted mean of the transition rates from the transient states to the absorbing state k.

The median lifetime from the transient state i to the absorbing state k is defined as the value of $t = \xi_i$ that satisfies $p_{ik}(\xi_i;\boldsymbol{z}) = 0.5$. The mean lifetime, $E_i = E\{T|X(0) = i\}$, is also a parameter of interest. Again, in terms of the transition probabilities, the mean lifetime is

$$E_i(\boldsymbol{z}) = \int_0^\infty S_i(t|\boldsymbol{z}) \ dt \ = \sum_{j=1}^{k-1}\sum_{v=1}^{k-1} a_{iv}(\boldsymbol{z})(-\frac{1}{\rho_v(\boldsymbol{z})})a^{vj}(\boldsymbol{z}),$$

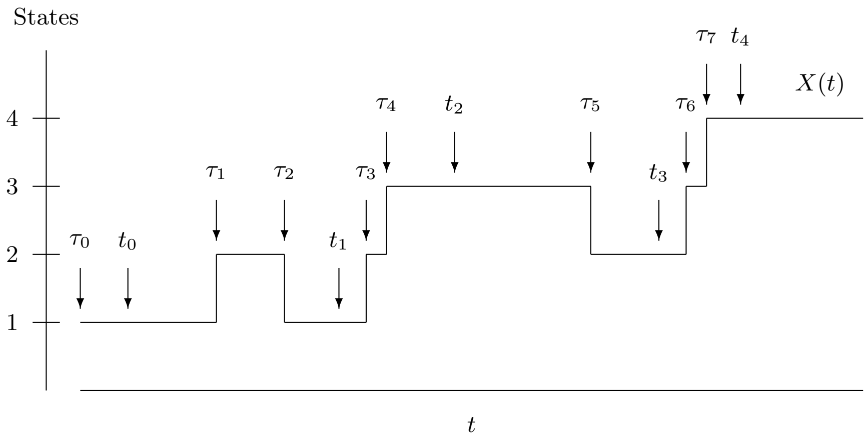
provided $\rho_v(\boldsymbol{z}) < 0, \ for every state v < k$. Finally, the residual mean lifetime, $m_i(t) = E\{T - t|X(0) = i\}$, can be expressed in terms of the Markov model as

$$m_i(t|\boldsymbol{z}) = \frac{\int_t^\infty S_i(u|\boldsymbol{z}) \ du}{S_i(t|\boldsymbol{z})} = \frac{\sum_{j=1}^{k-1}\sum_{v=1}^{k-1}\frac{1}{\rho_v(\boldsymbol{z})}a_{iv}(\boldsymbol{z})e^{\rho_v(\boldsymbol{z})t}a^{vj}(\boldsymbol{z})}{1 - p_{ik}(t;\boldsymbol{z})} \ .$$

The Data

The type of data collected will be different in each application and is directly dependent on the nature of the process. When the exact transition times of the process $\tau_1, \cdots, \tau_{m'}$ are available, see Figure 2, the statistical methods for estimating the parameters of the multi-state model are straightforward. Closed form solutions for the maximum likelihood estimates can be derived for

Fig. 2. An example of a process with 4 states where the τ_i are the actual transition times and where the t_i are the observation times of the process.



the basic model, and an approach similar to fitting exponential regression models can be used for the model with covariables.

In clinical studies in which each realization of the process is a different patient, it is very unusual to observe the exact transition times. The typical information available are the visits of the patients to the clinic, $t_0, t_1, \cdots, t_{m+1}$, as shown also in Figure 2. We assume that the data obtained on each subject are unequally spaced in time, and that the exact transition times are not available. For a model of $k = 4$ states, the following data correspond to weeks from the date of diagnosis and the state of the disease of the patient at that specific date (13).

Patient	*Data*
1	$(0,2)$ $(41,2)$ $(78,1)$ $(95,3)$ $(104,4)$
2	$(0,1)$ $(17,1)$ $(52,4)$
3	$(0,3)$ $(23,2)$ $(58,3)$ $(72,2)$

At the date of diagnosis, patient 2 was in state 1, and seventeen weeks later patient 2 was in state 1. This patient could have remained in state 1 for the whole 17 weeks, or could have transferred out of state 1 and back in again. Thirty-five weeks later, at week 52, the patient died. Survival times for these patients are 104 weeks for patient 1, 52 weeks for patient 2 and more than 72 weeks for patient 3. The data of patient 3 represent a censored observation.

The Likelihood Function

Suppose that the observation times of the process for a subject are $t_0 < t_1 < \cdots < t_m < t_{m+1}$, and that $x(t_0) = i_0, x(t_1) = i_1, \ldots, x(t_{m+1}) = i_{m+1}$ represent the observed states of the process at these particular times. Then the joint distribution of $X(t_1), X(t_2), \ldots, X(t_{m+1})$ given $X(t_0)$ and the vector of covariables $\mathbf{z}$ can be represented, using the Markov property (1) and conditional probabilities, as

$$\prod_{j=1}^{m+1} p_{i_{j-1},i_j}(\Delta t_j; \boldsymbol{z}), \tag{8}$$

where $\mathit{\Delta} t_j = t_j - t_{j-1}$.

The above expression represents the contribution to the likelihood function for this subject if the arrival time at the absorbing state is interval-censored, in other words, if t_{m+1} is not the known time of transition to the absorbing state. In survival studies, t_{m+1} may represent the exact arrival time at the absorbing state k, which is the lifetime, or it may represent the end of the study for this subject, which is the censuring time.

Let $T = \tau$ be the exact arrival time at k, and c be the censoring time for this subject. Then

$$t_{m+1} = \min(\tau, c) \;\text{ and }\; \delta = \begin{cases} 1 \text{ if } \tau \leq c \\ 0 \text{ if } \tau > c \,. \end{cases}$$

If $\delta = 1$, the contribution of this last transition to the likelihood is

$$f_{i_m}(\mathit{\Delta} t_{m+1}|\boldsymbol{z}) = \sum_{j=1}^{k-1} p_{i_m,j}(\mathit{\Delta} t_{m+1};\boldsymbol{z})\mu_j(\boldsymbol{z}) \;,$$

and if $\delta = 0$ the contribution is $S_{i_m}(\mathit{\Delta} t_{m+1}|\boldsymbol{z}) = 1 - p_{i_m,k}(\mathit{\Delta} t_{m+1};\boldsymbol{z})$. The above expression for $\delta = 1$ is a continuous time version of Kay's likelihood contribution (13).

The likelihood function for this subject can be then written as

$$L_h(\boldsymbol{\theta}) = \prod_{j=1}^{m} p_{i_{j-1},i_j}(\mathit{\Delta} t_j;\boldsymbol{z})\{f_{i_m}(\mathit{\Delta} t_{m+1}|\boldsymbol{z})\}^{\delta}\{S_{i_m}(\mathit{\Delta} t_{m+1}|\boldsymbol{z})\}^{1-\delta}. \tag{9}$$

The full likelihood can be obtained from the product of the individual contributions. The subject subscript h has been omitted for m, i_j, t_j, $\boldsymbol{z}$ and δ in the expression (9) and in the rest of this paper for simplicity.

Without the first term in expression (9) this likelihood function is equal to the likelihood for parametric models in survival analysis with censored observations. For a model with two states $k = 2$, and with constant transition intensities, this Markov model reduces to a survival analysis model using the exponential distribution. In particular $p_{11}(t|\boldsymbol{z}) = \exp\{-\mu(\boldsymbol{z})t\}$ and $p_{12}(t|\boldsymbol{z}) = 1 - \exp\{-\mu(\boldsymbol{z}t\}$, therefore the above contribution to the likelihood is $\{f(t_{m+1}|\boldsymbol{z})\}^{\delta}\{S(t_{m+1}|\boldsymbol{z})\}^{1-\delta}$.

The likelihood function (9) can be extended to the case of time-dependent covariables $\boldsymbol{z}(t)$ by replacing $\boldsymbol{z}$ by $\boldsymbol{z}_{j-1}$, where the covariables are asummed to be constant between two obsertations

$$\boldsymbol{z}(t) = \boldsymbol{z}_{j-1} \;\text{ for }\; t_{j-1} \leq t < t_j.$$

Parameter Estimation

Let $\boldsymbol{\theta}_{ij} = (\log\lambda_{ij}, \beta_{ij1}, \ldots, \beta_{ijp})$ be the parameters associated with the transition between states i to j, and $\boldsymbol{\theta}$ be a vector made up of the $\boldsymbol{\theta}_{ij}$ vectors. A log transformation is used to prevent the baseline transition intensities from taking negative values during the iterative process of estimation. Let $\eta_{ij} = \log q_{ij}(\boldsymbol{z})$ be the log transition intensity for a subject with covariables $\boldsymbol{z}$

$$\eta_{ij} = \log \lambda_{ij} + \beta_{ij1} z_1 + \cdots + \beta_{ijp} z_p.$$

Maximum likelihood estimates of $\boldsymbol{\theta}$ can be found by maximizing the log-likelihood function $l(\boldsymbol{\theta}) = \sum l_h(\boldsymbol{\theta})$ where

$$l_h(\boldsymbol{\theta}) = \sum_{j=1}^{m} \log\{p_{i_{j-1},i_j}(\varDelta t_j; \boldsymbol{z})\} + \delta \log\{f_{i_m}(\varDelta t_{m+1}|\boldsymbol{z})\} + (1-\delta)\log\{S_{i_m}(\varDelta t_{m+1}|\boldsymbol{z})\}.$$

The first derivative of the log-likelihood function with respect to θ_{uvl} is

$$\frac{dl_h(\boldsymbol{\theta})}{d\theta_{uvl}} = \left\{ \sum_{j=1}^{m} \frac{1}{p_{i_{j-1},i_j}(\varDelta t_j; \boldsymbol{z})} \frac{dp_{i_{j-1},i_j}(\varDelta t_j; \boldsymbol{z})}{d\eta_{uv}} + \frac{\delta}{f_{i_m}(\varDelta t_{m+1}|\boldsymbol{z})} \frac{df_{i_m}(\varDelta t_{m+1}; \boldsymbol{z})}{d\eta_{uv}} \right.$$
$$\left. + \frac{1-\delta}{S_{i_m}(\varDelta t_{m+1}|\boldsymbol{z})} \frac{dS_{i_m}(\varDelta t_{m+1}; \boldsymbol{z})}{d\eta_{uv}} \right\} \frac{d\eta_{uv}}{d\theta_{uvl}},$$

where the derivative of $f_i(t|\boldsymbol{z})$ with respect to η_{uv} is

$$\frac{df_i(t|\boldsymbol{z})}{d\eta_{uv}} = \sum_{j=1}^{k-1} \{ \frac{dp_{ij}(t; \boldsymbol{z})}{d\eta_{uv}} e^{\eta_{jk}} + p_{ij}(t; \boldsymbol{z}) \frac{de^{\eta_{jk}}}{d\eta_{uv}} \},$$

and where

$$\frac{dS_i(t|\boldsymbol{z})}{d\eta_{uv}} = -\frac{dp_{ik}(t; \boldsymbol{z})}{d\eta_{uv}}.$$

The derivative $dp_{ij}(t; \boldsymbol{z})/d\eta_{uv}$ in the three expressions above is

$$\frac{dp_{ij}(t; \boldsymbol{z})}{d\eta_{uv}} = \sum_{r=1}^{k} \sum_{s=1}^{k} a_{ir}(\boldsymbol{z}) w_{rs}^{uv}(t; \boldsymbol{z}) a^{sj}(\boldsymbol{z}) \tag{10}$$

where

$$w_{rs}^{uv}(t; \boldsymbol{z}) = \begin{cases} g_{rs}^{uv}(\boldsymbol{z})(e^{\rho_r(\boldsymbol{z})t} - e^{\rho_s(\boldsymbol{z})t})/(\rho_r(\boldsymbol{z}) - \rho_s(\boldsymbol{z})) & \text{if } r \neq s \\ g_{rr}^{uv}(\boldsymbol{z}) t e^{\rho_r(\boldsymbol{z})t} & \text{if } r = s \end{cases},$$

and $g_{rs}^{uv}(\boldsymbol{z})$ is the (r,s) entry in

$$\mathbf{G}^{uv}(\boldsymbol{z}) = \mathbf{A}(\boldsymbol{z})^{-1} \frac{d\mathbf{Q}(\boldsymbol{z})}{d\eta_{uv}} \mathbf{A}(\boldsymbol{z}) \ .$$

The derivative of $d\eta_{uv}/d\theta_{uvl}$ is

$$\frac{d\eta_{uv}}{d\theta_{uvl}} = \begin{cases} 1 & \text{for } \ l = 1 \\ z_{l-1} & \text{for } \ l = 2, \ldots, p+1. \end{cases}$$

The subscripts u and v refer to the transition between states u to v.

Quasi-Newton algorithms can be used to minimize $-2l(\boldsymbol{\theta})$ using only the likelihood function and finite differences to obtain numerical approximations of the derivatives, or by using the likelihood function and explicit expressions for the derivatives. A discussion of these two approaches can be found in Dennis and Schnabel (6) as well as a modular system of algorithms for unconstrained minimization. Dennis and Schnabel (6) also provide algorithms for computing numerical approximations of the second derivatives of the log–likelihood using finite differences of the original function or the gradients if they are available.

Once we have the maximum likelihood estimates of the parameters of the transition intensity matrix $\mathbf{Q}(\boldsymbol{z})$, we also have estimates of the transition probability matrix $\mathbf{P}(t;\boldsymbol{z};\boldsymbol{\theta})$. In particular, an estimates of $p_{ij}(t;\boldsymbol{z};\boldsymbol{\theta})$ can be obtained as $p_{ij}(t;\boldsymbol{z};\hat{\boldsymbol{\theta}})$.

An estimate of the asymptotic covariance matrix of $\hat{\boldsymbol{\theta}}$ is obtained by inverting the negative of the empirical information matrix,

$$\widehat{\mathbf{V}}(\hat{\boldsymbol{\theta}}) = -\left\{\frac{d^2L(\boldsymbol{\theta})}{d\boldsymbol{\theta}d\boldsymbol{\theta}'}\right\}^{-1}_{\theta=\hat{\theta}} .$$

The estimate of the asymptotic variance of $p_{ij}(t;\boldsymbol{z};\hat{\boldsymbol{\theta}})$ can be found using the δ method as

$$\widehat{V}\{p_{ij}(t;\boldsymbol{z};\hat{\boldsymbol{\theta}})\} = \left\{\frac{dp_{ij}(t;\boldsymbol{z};\boldsymbol{\theta})}{d\boldsymbol{\theta}}\right\}'_{\theta=\hat{\theta}} \widehat{\mathbf{V}}(\hat{\boldsymbol{\theta}}) \left\{\frac{dp_{ij}(t;\boldsymbol{z};\boldsymbol{\theta})}{d\boldsymbol{\theta}}\right\}_{\theta=\hat{\theta}}$$

where $dp_{ij}(t;\boldsymbol{z};\boldsymbol{\theta})/d\boldsymbol{\theta}$ can be evaluated using expression (10). An estimate of the survival function can be obtained directly from the estimate of the transition probability matrix as $S_i(t|\boldsymbol{z}) = 1 - p_{ik}(t;\boldsymbol{z};\hat{\boldsymbol{\theta}})$. An approximate variance for $\widehat{S}(t|\boldsymbol{z})$ is obtained as

$$\widehat{V}\{\hat{S}(t|\boldsymbol{z})\} = \widehat{V}\{p_{ik}(t;\boldsymbol{z};\hat{\boldsymbol{\theta}})\}.$$

The Natural Course of Diabetic Retinopathy

Diabetic retinopathy currently is the leading cause of new cases of blindness in people aged 20 to 74 years in the United States, and is considered a progressive disease among people with insulin-dependent (type I) diabetes mellitus (IDDM).

Improvement of early stages of retinopathy as part of the natural course has been poorly understood. In the past, physicians believed diabetic retinopathy was a strictly progressive disease. Using a basic multi-state Markov model Garg, Marshall, Chase, et al. (1990) have shown that the natural course of early diabetic retinopathy involves both progression and regression.

The natural course of early diabetic retinopathy in young subjects with type I diabetes was evaluated during 693 patient visits for 259 subjects over a mean of 2.4 years. All 259 subjects had direct ophthalmoscopy (with pupils dilated) by at least two examiners (one ophthalmologic and one pediatric), followed by color retinal photography, intravenous fluorescein angiography , and slit-lamp examinations. The retinal specialist graded retinal findings with a modified Airlie House classification of diabetic retinopathy. A grade of I indicates no retinopathy; grades II–III, microaneurysms or microaneurysms and one other finding; grades IV-V, advanced background changes with intra retinal microvascular abnormalities; and grade VI, proliferative retinopathy. The category assigned was that of the more severely involved eye.

Based on this classification, a 4–state Markov model was used considering grades I, II–III and IV–V as transient states and grade VI as an absorbing state, as shown in Figure 3. In this case the exact arrival times at the absorbing state were interval-censored and the likelihood function (8) was used to estimate the paramaters.

Fig. 3. A model with covariables for diabetic retinopathy

$$\begin{tikzcd}
\text{Grade I} \arrow[r, shift left, "\lambda_{12}(\mathbf{z})"] & \text{Grades II–III} \arrow[l, shift left, "\lambda_{21}(\mathbf{z})"] \arrow[r, shift left, "\lambda_{23}(\mathbf{z})"] & \text{Grades IV–V} \arrow[l, shift left, "\lambda_{32}(\mathbf{z})"] \arrow[r, "\mu_3"] & \text{Grade VI}
\end{tikzcd}$$

The influence of duration of diabetes, HbA1, age and sex on the transition intensities between various stages of diabetic retinopathy was evaluated. A model was fit for each of these four factors for all transitions except the transition from state 3 to the absorbing state, as shown in Figure 3. Duration of diabetes, HbA1 and age are time dependent covariables, but they were very stable between two observed times.

Table 1. Summary of models fit

N	Model	Parameters	$-2LogLike$	χ^2	P–value	Ref	AIC
1	Null	5	602.66	–	–	–	612.66
2	HbA1	9	592.85	9.81	0.044	1	610.85
3	Age	9	586.90	15.76	0.003	1	604.90
4	Duration	9	580.93	21.73	0.001	1	598.93
5	Sex	9	599.21	3.45	0.486	1	617.21
6	HbA1+			14.84	0.005	4	
	Duration	13	566.09	26.76	0.001	2	592.09*
7	Age+			5.82	0.213	4	
	Duration	13	575.11	11.79	0.019	3	601.11

* Best model

Table 1 shows a summary of the models and the significance of the effect of each factor. Age is a significant factor in the development of diabetic retinopathy, but its effect is partially due to the confounding effect of duration of diabetes. Model 7 shows that the effect of age is no longer significant when duration of diabetes is also in the model. Sex shows no significant effect on the process of early diabetic retinopathy. Duration and HbA1 are the two most important independent factors affecting the process of progression and regression of early diabetic retinopathy.

Table 2 shows parameter estimates for a model which includes duration of diabetes and HbA1 as covariables. This model involves the only 3 regression parameters that are statistically significant in both the univariate and multivariate models 2, 4 and 6 for these two variables.

From this table it can be inferred that one unit of the absolute variation in HbA1 represents an increase of 35.6% on the transition rate from normal to grades II-III in early diabetes retinopathy.

Table 2. Maximum likelihood estimates for the final model

Parameter	Estimate	Std. Error	Relative Risk
λ_{12}	0.0003	0.0004	–
λ_{21}	0.0173	0.0036	–
λ_{23}	0.0134	0.0036	–
λ_{32}	2.0540	2.0570	–
μ_3	0.0080	0.0047	–
$\beta_{12,HbA1}$	0.3044	0.1010	1.356
$\beta_{12,Duration}$	0.1386	0.0453	1.149
$\beta_{32,Duration}$	-0.2900	0.0825	0.748

One additional year in duration of diabetes represents an increase of 15% in the transition rate from normal to grades II-III and a reduction of 25% for the chances of regression from grades IV–V to grades II-III.

4 Non-Homogeneous Markov Models

In the course of modelling a multi-state Markov process it is natural to raise the question of whether the transition intensities are constant over time. In some processes the transition parameters will be clearly time-dependent. The homogeneous model can be extended to the non-homogeneous case by assuming that the multi-state model has time-varying parameters.

Different parametric and semi-parametric functions of the time can be proposed to model the transition intensities. From the wide selection of parametric families of hazard functions one of the most popular is the Weibull hazard function, $\lambda(t) = p\lambda(\lambda t)^{p-1}$. This is often used in survival analysis, and the exponential hazard function ($p = 1$) is a special case. Splines and other smoothers are attractive semi-parametric approaches for describing the changes in the transition intensities over time. If the transition intensities are arbitrary functions of time, the Kolmogorov system of differential equations does not have a closed form solution and it becomes increasingly more difficult, if not impossible, to obtain a simple expression for the likelihood function in contrast with the homogeneous model. As was mentioned before, when the process has unidirectional transitions, a simple closed form solution for the transition probability matrix can be found by systematic integration.

A practical solution of this problem is to subdivide the period of follow-up into K intervals $[\tau_1, \tau_2), [\tau_2, \tau_3), \ldots, [\tau_K, \infty)$, and assume a constant intensity matrix in each interval. The intensity matrix can be written as $\mathbf{Q}(t) = \mathbf{Q}_l$, where $t \in [\tau_l, \tau_{l+1})$. A local solution of the Kolmogorov differential equations can then be found for each interval and a general solution can be obtained applying the Chapman-Kolmogorov equation as in expression (5).

Parametric forms for the transition intensities for this piecewise constant model are,

$$\lambda_{ijl} = \lambda_{ij} t_l^{\gamma_{ij}} \quad \text{and} \quad \mu_{il} = \mu_i t_l^{\gamma_{ik}}$$

or

$$\lambda_{ijl} = \lambda_{ij} e^{\gamma_{ij} t_l} \quad \text{and} \quad \mu_{il} = \mu_i e^{\gamma_{ik} t_l},$$

where t_l is the center point in the interval $[\tau_l, \tau_{l+1})$, and where λ_{ijl} and μ_{il} are the elements (i,j) and (i,k) of the intensity matrix $\mathbf{Q}_l$. The advantage of assuming a parametric form for the transition intensities is the dramatic reduction in the number of parameters involved in the model. In addition, a test for a constant transition intensity from i to j is equivalent to the test $H_0 : \gamma_{ij} = 0$. A global test for homogeneity of the process can be performed by doing a simultaneous test for all the γ's.

The joint distribution of $X(t_1), X(t_2), \ldots, X(t_{m+1})$ given $X(t_0)$ can be represented as

$$\prod_{j=1}^{m+1} p_{i_{j-1},i_j}(t_{j-1}, t_j)$$

where $p_{ij}(s,t)$ is the element (i,j) of the resulting product of matrices in (5). Using the same argument as in section (3.5), this expression represents the contribution to the likelihood for this subject if the arrival time at the absorbing state is interval-censored. If t_{m+1} is the lifetime for this subject or the censuring time, then the contribution to the likelihood is equivalent to expression (9) and can be written as

$$L(\boldsymbol{\theta}) = \prod_{j=1}^{m} p_{i_{j-1},i_j}(t_{j-1}, t_j)\{f_{i_m}(t_m, t_{m+1})\}^{\delta}\{S_{i_m}(t_m, t_{m+1})\}^{1-\delta},$$

where

$$f_{i_m}(t_m, t_{m+1}) = \sum_{j=1}^{k-1} p_{i_m,j}(t_m, t_{m+1})\mu_j(t_{m+1})$$

and where $S_{i_m}(t_m, t_{m+1}) = 1 - p_{i_m,k}(t_m, t_{m+1})$. Quasi-Newton algorithms can be used to minimize $-2l(\boldsymbol{\theta})$ using only the likelihood function and finite differences to obtain approximations for the first two derivatives (Dennis and Schnabel (6)).

This model can be easily extended to include time-dependent covariables. The model for the transition intensities can be written as

$$\lambda_{ij}(t, \mathbf{z}(t)) = \lambda_{ij}(t)e^{\boldsymbol{\beta}_{ij}\mathbf{z}(t)} \text{ and } \mu_i(t, \boldsymbol{z}(t)) = \mu_i e^{\boldsymbol{\beta}_{ik}\boldsymbol{z}(t)}$$

where $\lambda_{ij}(t)$ is the baseline transition intensity from the state i to the state j, $\boldsymbol{\beta}_{ij}$ is a vector of regression coefficient, and $\mathbf{z}(t) = \mathbf{z}_l$ is the vector of covariables, constant for every value of t in $[\tau_l, \tau_{l+1})$. The baseline transition intensities $\lambda_{ij}(t)$ can be modelled using a parametric form. In that case the model can be written

$$\lambda_{ijl} = \lambda_{ij}t_l^{\gamma_{ij}}e^{\boldsymbol{\beta}_{ij}\boldsymbol{z}_l} \text{ and } \mu_{il} = \mu_i t_l^{\gamma_{ik}}e^{\boldsymbol{\beta}_{ik}\boldsymbol{z}_l}.$$

Bibliography

[1] Chiang, C.L. (1968). Introduction to Stochastic Processes in Biostatistics. Wiley, New York.

[2] Cox, D.R. and Miller, H.D. (1965). The Theory of Stochastic Processes. Methuen, London.

[3] Cox, D.R. (1972). Regression models and life-tables (with discussion). Journal of the Royal Statistical Society, Series B **34**, 187-220.

[4] Cox, D.R. and Oakes, D. (1984). Analysis of Survival Data. Chapman and Hall, London.

[5] Davies, G.S. (1985) A Note on a Continuous-Time Markov Manpower Model. Journal of Applied Probability, **22**, 932–938.

[6] Dennis, J.E., Jr., and Schnabel, R.B. (1983). Numerical Methods for Unconstrained Optimization and Nonlinear Equations. Prentice-Hall, Englewood Cliffs, New Jersey.

[7] Gantmacher, F.R. (1959). The Theory of Matrices, vols 1–2, Chelsea New York.

[8] Garg, S.K., Marshall, G., Chase, H.P., Jackson, W., Archer, P., Crews, M. (1990). The Use of the Markov Processes in Describing the Natural Course of Diabetic Retinopathy. Archives of Ophthalmology, **108**, 1245–1247.

[9] Hochstadt, H. (1963). Differential Equations: A Modern Approach. Dover, New York.

[10] Horn, R.A., Johnson, C.R. (1985). Matrix Analysis, Cambridge University Press, Cambridge.

[11] Kalbfleisch, J.D., Lawless, J.F. and Vollmer, W.M. (1983). Estimation in Markov Models from Aggregate Data. Biometrics **39**, 907–919.

[12] Kalbfleisch, J.D. and Lawless, J.F. (1985). The Analysis of Panel Data Under a Markov Assumption. Journal of the American Statistical Association **80**, 832–871.

[13] Kay, R. (1986). A Markov Model for Analysing Cancer Markers and Disease States in Survival Studies. Biometrics **42**, 855–865.

[14] Klein J.P., Klotz, J.H. and Grever, M.R. (1984). A Biological Marker Model for Predicting Disease Transitions. Biometrics **40** 927–936.

[15] Longini Jr. I.M., Clark, W.S., Byers, R.H., Ward, J.W., Darrow, W.W., Lemp, G.F. and Hethcote, H.W. (1989). Statistical Analysis of the Stages of HIV Infection Using a Markov Model. Statistics in Medicine **8**, 831–843.

[16] Madsen, H., Spliid, H., Thyregod, P. (1985). Markov Models in Discrete and Continuous Time for Hourly Observations of Cloud Cover. Journal of Climate and Applied Meteorology **24**, 629–639.

[17] Raman, S. and Chiang, C.L. (1973). On a Solution of the Migration Process and the Application to a Problem in Epidemiology. Journal of Applied Probability **10**, 718–727.

[18] Schoen, R. and Land, K.C. (1979). A General Algorithm for Estimating a Markov-Generated Increment-Decrement Life Table With Applications to Marital-Status Patterns. Journal of the American Statistical Association **74**, 761–776.

[19] Wilkinson, J.H. (1965). The Algebraic Eigenvalue Problem. Clarendon Press, Oxford.

Article 1.2

Multi-State Models and Diabetic Retinopathy

Guillermo Marshall and Richard H. Jones

Pontificia Universidad Católica de Chile and
University of Colorado HSC, Denver, U.S.A.

Abstract. This paper discusses the application of a multi-state model to diabetic retinopathy under the assumption that a continuous time Markov process determines the transition times between disease stages. The multi-state model consists of three transient states that represent the early stages of retinopathy, and one final absorbing state that represent the irreversible stage of retinopathy. By using a model with covariables, we explore the effects of factors that influence the onset, progression, and regression of diabetic retinopathy among subjects with insulin-dependent diabetes mellitus. We can also introduced time-dependent covariables in the model by assuming that the covariables remain constant between two observations. We can also obtained survival-type curves from each stage of the disease and for any combination of patient risk factors.

Keywords: Markov models; Insulin-Dependent Diabetes Mellitus; Survival Functions; Longitudinal Data. [1]

1 Introduction

The classification of early diabetic retinopathy on a scale from Grade I to Grade VI according to the modified Airlie House classification (1; 2) suggests that multi-state modelling might offer an innovative methodology to analyze the natural course of this disease and may be the most appropriate methodology for finding the factors that influence this disease process. Such analysis will likely not only assess more accurately the effects of the risk factors in the disease process, but will also allow the prediction of transition times between disease stages.

Previous studies (3; 4; 5; 6) have used contingency tables and logistic regression models to find patient risk factors associated with progression of diabetic retinopathy. Only one previous study (7) has modelled the effects of risk factors consistently with the longitudinal nature of the disease process using proportional hazard models. No previous study, however, has modelled diabetic complications using a multi-state model that allows progression and regression transitions among the different stages of diabetic retinopathy.

A multi-state Markov model without covariates has had successful applications to the stages of cancer (8), the stages of HIV infection (9), and the stages of diabetic retinopathy (10) among chronic diseases. In all of these cases the major problem is the type of data collected from the respective longitudinal medical studies. Ideally, researchers would like to observe every transition

[1] Marshall G and Jones RH. (1995) Multi-state models and diabetic retinopathy. Statistics in Medicine, 14 (18), 1975-1983

time in a patient's disease process. In general, however, one can only collect observations on stage of the process at the time of the patients' irregular clinical visits.

Marshall (11) and Marshall and Jones (12) proposed the extension of this model in various directions. One such direction is the inclusion of covariates in the model. By introducing covariates into the models, one can not only describe the natural course of the disease, but also find the factors associated with progression and regression between disease stages. Marshall and Jones (13) have developed a computer program called MARKOV to fit a general k-state Markov model.

2 Data

Two hundred and seventy-seven subjects who had Type I diabetes for at least five years had a mean age of 18 years and ranged in age from 14 to 29 years when initially seen at the Eye-Kidney Clinic of the Barbara Davis Center for Childhood Diabetes at the University of Colorado Health Sciences Center. The Eye-Kidney Clinic is open to all patients 14 years of age or older, and who have had Type I diabetes for at least three years.

The average duration of insulin dependent diabetes mellitus for this population is approximately 10 years, ranging from three to 28 years. The gender distribution is uniform. In data collection for this study, all subjects were seen longitudinally at least twice with visits at an average of one year apart for a mean follow-up of three years. A total of 882 patient visits occurred during the study period.

At each visit, a retinal specialist graded retinal findings using a modified Airlie House classification (1; 2) in which Grade I indicates no retinopathy; Grade II indicates microaneurysms only; Grades III and IV indicate intermediate stages of background retinopathy and grades V and VI indicate preproliferative and proliferative retinopathy, respectively. The worse eye grade for each visit was used to define the subject's state at the time of the visit.

3 The Multi-state Markov Model

The four-state Markov model that we consider for the analysis of these data includes three transient disease states: Grade I, Grades II-III and Grades IV-V of early retinopathy ($j = 1, 2, 3$), and one absorbing state 4 representing retinopathy or Grade VI. In this model the transient states are ordered according to j, and instantaneous transition, represented by the intensities λ, can ocurr from state j to the adjoining states $j-1$ or $j+1$ as shown in Figure 1. No direct transitions are allowed from an early stage of retinopathy to the absorbing state (except from the state IV-V), and if transitions like this occurred, the model assumes that unobserved transitions have occurred before the final transition. Alternatively, we might have considered a six-state Markov using the six grades of retinopathy. However, in addition to the attractiveness of a model with a reduced number of parameters, the four-state model can reduce substantially the false transitions due to misclassification.

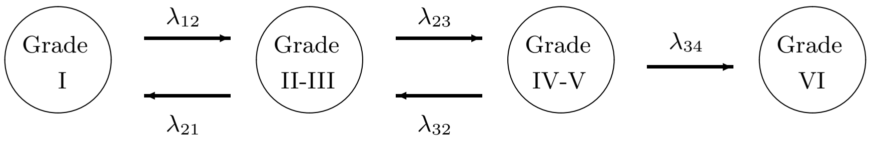


Fig. 1. The Multistate model for four stages of diabetic retinopathy defined on the basis of eye grades according to the Airlie House classification.

Assuming that the underlying process is a Markov process, we represent this model using the transition intensity matrix $\mathbf{\Lambda}$ as

$$\mathbf{\Lambda} = \begin{pmatrix} -\lambda_{12} & \lambda_{12} & 0 & 0 \\ \lambda_{21} & -(\lambda_{21}+\lambda_{23}) & \lambda_{23} & 0 \\ 0 & \lambda_{32} & -(\lambda_{32}+\lambda_{34}) & \lambda_{34} \\ 0 & 0 & 0 & 0 \end{pmatrix}. \tag{1}$$

We establish the relation between the transition probability matrix $\mathbf{P(t)}$ and the transition intensity matrix $\mathbf{\Lambda}$ with the Kolmogorov forward differential equations

$$\frac{d\mathbf{P}(t)}{dt} = \mathbf{P}(t)\mathbf{\Lambda}(t) \tag{2}$$

where the element (i, j) of the matrix $\mathbf{P(t)}$ represents the probability of a transition from the state i to the state j in a time interval t, denoted as $p_{ij}(t)$. We can express the solution to this system of differential equations as

$$\mathbf{P}(t) \; = \; \mathbf{A} diag\{ \; e^{\rho_1 t}, e^{\rho_2 t}, \ldots, e^{\rho_k t} \; \}\mathbf{A}^{-1} \tag{3}$$

where $\mathbf{A}$ is the square matrix containing in column i the eigenvector associated with the eigenvalue ρ_i of the transition intensity matrix $\mathbf{\Lambda}$. For a more detailed discussion about Markov processes, see Cox and Miller (14).

We can extend the model by introducing covariables as a proportional factor in the baseline transition intensities λ's. We represent the regression for the element (i, j) of the transition intensity matrix as $\mathbf{\Lambda}$

$$\lambda_{ij}(\mathbf{z}) = \lambda_{ij} e^{\beta'_{ij}\mathbf{z}} \tag{4}$$

where β_{ij} is the vector of regression coefficients associated with the vector of covariables $\mathbf{z}$ for the transition between the states i and j. Note that model 4 for the transition intensity $\lambda_{ij}(\mathbf{z})$ resembles the proportional hazard model with constant hazard function. We can use the resulting transition intensity matrix $\mathbf{\Lambda}(\mathbf{z})$ for a subject with vector of covariates $\mathbf{z}$ in equations (2) and (3) to compute the transition probability matrix $\mathbf{P(t, z)}$. The elements $p_{ij}(t, \mathbf{z})$'s of this transition probability matrix constitute the contribution of each observation to the likelihood function.

4 Model Selection

We must consider two types of model selections procedures in the context of this multi-state Markov model. The first, more classical in statistical analysis, is the selection of covariates associated significantly with the progression and regression of the process. Given the large number of parameters associated with each covariate in model (4), it seems reasonable to consider a forward selection procedure. The second, more specific to this Markov model, relates to the selection of the most parsimonious representation of the association between each covariate and the disease process.

Consider the case of a model with a single covariate. In the context of this four-state model for diabetic retinopathy, there are three natural models for representing the effect of the covariate in the progression and regression of the disease process. The first, named the saturated model, is defined as the model in which the effect of the covariate differs in each of the five disease transitions (Figure 1). In this model we have a total of 10 parameters, five baseline transition intensities, and five different regression coefficients. The second model, named the progression and regression (PR) model, is defined as the model in which the effect of the covariate is the same for all progression transitions, and the same for all regression transitions. More formally, we formulate this model by assuming that the null hypothesis $H_0' : \beta_{j,j+1} = \beta_p, j = 1,2,3$ and $\beta_{j,j-1} = \beta_r, j = 2,3$ Under this hypothesis the number of parameters associated with each covariable reduces from five to only two. We can write the proportional intensity model (4) under this hypothesis as

$$\lambda_{ij}\left(z\right) = \begin{cases} \lambda_{ij} e^{\beta'_p \mathbf{z}} \; j = i+1 \\ \lambda_{ij} e^{\beta'_r \mathbf{z}} \; j = i-1 \end{cases} \tag{5}$$

Finally, the third model, called the progression minus regression (PMR) model, is defined as the model in which the effect of the covariate is the same for all progression transitions and the same, but with a sign change, for all regression transitions. Formally, we can formulated the hypothesis as $H_0'' : \beta_p = \beta_r = \beta$, provided that H_0'' is true, and therefore model (4) reduces to

$$\lambda_{ij}\left(z\right) = \begin{cases} \;\lambda_{ij} e^{\beta' \mathbf{z}} \;\; j = i+1 \\ \lambda_{ij} e^{-\beta' \mathbf{z}} \; j = i-1 \end{cases} \tag{6}$$

Assuming that this hypothesis is true, we reduce by four the number of parameters associated with this covariate with respect to model (4) and by one parameter with respect to model (5). This reduction becomes extremely important when we include more variables in the model. In addition to the reduction in the number of parameters, if the null hypotheses H_0' and (or) H_0'' are true we can expect, according to our experience, to have a more robust estimation and substantially more power to assess the effect of the covariate in the disease process, and less chance that we overfit the model. We can use the likelihood ratio and the Wald tests to test these two hypotheses. The first is more convenient for covariate selection, while the Wald test is more convenient for testing H_0' and (or) H_0'' , since we need only fit the saturated model.

5 Estimation of Parameters

The major distinction of this multi-state Markov model with respect to other related techniques is its ability to analyze unobserved transition times based on the observation of the process at arbitrary times. Typical information collected at each visit from the patient includes the grade of diabetic retinopathy and other disease-related measurements. If i and j represent the observed states of the process at times s and t respectively, then the contribution of this observed transition to the likelihood function is $p_{ij}(t-s;\mathbf{z})$, that is, the element (i,j) of the transition probability matrix (3) evaluated at time $t-s$ and with covariate $\mathbf{z}$.

The total contribution of an individual to the likelihood function is the result of the product of the contribution from each observed transition. The full likelihood function is the product of all individual contributions. The model can be adapted to handle time- dependent covariables by replacing the time-invariate covariate contribution, $p_{ij}(t-s;\mathbf{z})$, with $p_{ij}(t-s;\mathbf{z}(s))$ by assuming that the time-dependent covariate remains constant between the two consecutive times s and t. Note that the times s and t are more often arbitrary times and they do not necessarily represent the actual transition times of the underlying disease process. Furthermore, given the form in which the data is collected, we must assume that more than one transition may possibly occur between these two observed times.

Maximum likelihood estimates for λ's and β's can be obtained by maximizing the likelihood function with respect to these parameters, and asymptotic estimates of the standard errors of the estimates can be obtained by inverting the empirical information matrix. Quasi-Newton algorithms can be used to find the maximum likelihood estimates using only an analytical expression for the likelihood function and using finite differences to obtain numerical approximations of the derivatives. Given the high cost of the evaluation of the likelihood function in this case, this algorithm can be significantly accelerated by using an analytic expression for the first derivatives. In both situations the second derivative is updated at each iteration by using Cholesky or QR factorization. A complete discussion of these methods can be found in Dennis and Schnabel (15).

6 Survival Curves

This type of data can also be analyzed using more traditional techniques found in survival analysis. If we denote T as the random variable representing the time free of state 4 (Grade VI) retinopathy, we can use Cox's regression model to find factors associated with the distribution of T. The problem with using this model and other classical survival analysis models is the high percentage, 98% in this case, of right censoring in the data. On the other hand, an important amount of data representing transitions between intermediate stages of the disease process is collected during the study period. This data contains valuable information about the disease process and can be used to find the various factors that are associated with the progression and regression of the various stages of the disease.

The multi-state model can be seen as a natural generalization of classical survival analysis models. Instead of having one transient and one absorbing state that characterize survival analysis, the multi-state model allows multiple transient states and the same absorbing final state. This characteristic can make this model an approach significantly more efficient in analyzing highly censored data. This is particularly true when most of the transition data are observed between intermediate states, such as in diabetic retinopathy. The functional relationship between the survival function and the transition probability matrix can be obtained by the equation

$$S_i(t|\mathbf{z}) = 1 - p_{i4}(t;\mathbf{z})$$

where $S_i(t|\mathbf{z})$ is the survival function from the state i for a subject with covariables $\mathbf{z}$, and where $p_{ik}(t;\mathbf{z})$ is the element (i,k) of the transition probability matrix $\mathbf{P(t;z)}$. Although the transition intensities are time-invariant, the associated hazard function is

$$h_i(t|\mathbf{z}) = \frac{-dS_i(t|\mathbf{z})/dt}{S_i(t|\mathbf{z})} = \frac{\sum_{j=1}^{4} p_{ij}(t;\mathbf{z})\lambda_{j4}(\mathbf{z})}{1 - p_{i4}(t;\mathbf{z})} = \frac{p_{i3}(t;\mathbf{z})}{1 - p_{i4}(t;\mathbf{z})}\lambda_{34}(\mathbf{z})$$

since $\lambda_{14}(\mathbf{z}) = \lambda_{24}(\mathbf{z}) = \lambda_{44}(\mathbf{z}) = 0$, and where the expression in the numerator is direct consequence of the Kolmogorov forward differential equation (2).

7 Results

At the beginning of the study, the patients were distributed among the four stages of diabetic retinopathy as 42%, 53%, 5%, and 0%, respectively. By definition in this study, stage 4 started with no subjects. The distribution at the end of the study period was 26%, 53%, 19%, and 2%, respectively. Note that these probabilities distributions do not correspond to a fixed period of time for each subject, so they are not valid information for estimating transition probabilities.

A single-covariate Markov model was used to assess the individual effects of factors associated with diabetic retinopathy using a custom-designed computer program8. The full model with five regression coefficients, model (4) , the progression and regression model with two regression coefficients, model (5) , and the progression minus regression model with only one regression coefficient, model (6) , were fitted to each factor considered in this study (Table 1). For each covariate, the most parsimonious model among these three was found using the likelihood ratio test. If a factor based on the best model was found to be significantly associated with the disease process, the parsimonius representation of this factor was later used for multiple regression analysis.

The duration of diabetes, the age of the subject, and the mean HbA1c levels (mean of all assessments at or before visit time) were the factors most associated with transitions of diabetic retinopathy. Diastolic and systolic blood pressure and values of HbA1c at visit times were also associated with the disease process. All other factors, including gender, mean cholesterol level levels (mean of all assessments at or before visit time), family history of hypertension, systolic blood pressure, and a history of smoking, were not significantly associated with changes in diabetic retinopathy. The significance of the association between these factors and transition times was tested using the likelihood ratio test (Table 1). The only three factors in this study that are time-independent covariates are gender, family history of hypertension, and a history of smoking.

Duration of diabetes shows similar effects in all progressive transitions and similar effects in all regressive transitions. Model (5) is chosen as the best representation for the association of

Table 1. Likelihood ratio test of single-covariable Markov models for various factors associated with diabetic retinopathy using the full model (4), the PR model (5), and the PMR model (6). All tests are compared to a basic model without covariates.

Factor	Full Model	PR Model	PMR Model
	$\chi^2(5)$	$\chi^2(2)$	$\chi^2(1)$
Duration of Diabetes	58.2	54.7*	47.5
Age	33.5*	26.3	20.2
Mean HbA$_{1c}$	27.2	22.2	22.2*
Diastolic Blood Pressure	12.0	10.9	10.5*
HbA$_{1c}$ at the Visit	10.7	9.0	8.9*
Gender	9.8*	3.6	1.6
Smoking	9.4	4.1	3.6*
Systolic Blood Pressure	6.7	6.4	6.1*
Cholesterol	5.0	4.5	4.4*
Family Hx Hypertension	4.5	4.3*	0.9

* Best model based on LR test

this factor and diabetic retinopathy. The regression coefficient estimates for this model were $\beta = (0.0528, -0.2223)$, showing a significant departure from the assumption of model (6). Based on the standard errors of the estimates, (0.02774,0.0456), and their correlation coefficient, r=0.5295 , we can construct a Wald test for the hypothesis $H_0'' : \beta_p = \beta_r = \beta$, associated with model (6) . By using $\mathbf{L} = (1, 1)'$, the Wald statistic is

$$W = (\mathbf{L}'\widehat{\beta})'(\mathbf{L}'\widehat{\mathbf{V}}_{\widehat{\beta}}\mathbf{L})^{-1}(\mathbf{L}'\widehat{\beta}) = \frac{0.0288}{0.0042} = 6.80 \tag{7}$$

This value has an associated p-value lower than 0.01 based on the chi-square distribution with one degree of freedom. The equivalent likelihood ratio test for this hypothesis is $-2\log\{L_6/L_5\} = 54.7 - 47.5 = 7.2$ (Table 1). These two results confirm that the PMR model does not hold for duration of diabetes. Confidence intervals for the parameters in the model can be obtained by using a Wald-type test based on normal approximation.

Table 2. Parameter estimates and standard errors for the final multiple regression model.

Factor	Parameter	Estimate	Standard Error
Baseline	λ_{12}	0.0566	0.0075
Baseline	λ_{21}	0.0121	0.0024
Baseline	λ_{23}	0.0163	0.0035
Baseline	λ_{32}	0.0746	0.0243
Baseline	λ_{34}	0.0024	0.0011
Duration of Diabetes	β_{p1}	0.0729	0.0283
Duration of Diabetes	β_{r1}	-0.2084	0.0461
Mean HbA$_{1c}$	$\beta_{p2} = -\beta_{r2}$	0.2128	0.0386
Diastolic Blood Pressure	$\beta_{p3} = -\beta_{r3}$	0.0178	0.0056

Table 2 gives the estimates and the standard errors of the estimates for the parameters of the final multiple regression model. Duration of diabetes remained the most important factor for explaining changes in diabetic retinopathy. As expected, cumulative HBA1c was the second most important clinical variable associated with transitions in retinopathy. The additional contribution of this factor in terms of the likelihood ratio chi-square test is slightly superior to the chi-square obtained without controlling for duration of diabetes. Diastolic blood pressure also remained in the model showing that it is an independent factor associated with diabetic retinopathy.

The baseline parameters represent the transition rates from one stage to another for a subject with average risk factors (in our study these numbers are: 10.7 years of duration of diabetes, a HbA1c value of 11.8%, and a value of diastolic blood pressure of 70) for a given period of time, in this study one month. By multiplying the baseline transition estimate from stage 3 to stage 4 for 12 month and 100 subjects, we conclude that that an average of $2.88 (= 0.0024 \times 12 \times 100)$ transitions will occurr from stage 3 to stage 4 in a period of one year in a group of subjects will average risk factors. Similar conclusions can be made from the remaining baseline transition estimates. The parameters associated with the covariates can be interpreted similarly to the regression coefficients in the Cox regression model. The increment of one year of duration of diabetes will increase the risk of progression on the disease process 7.5% (e0.0729=1.075) and reduces the chances of regression in the disease process 19% (e-0.2084=0.81).

Figure 2 shows estimated survival curves of the probability of remaining free of state 4 (Grade VI) retinopathy for a subject with eight years since the onset of diabetes, 12% of HbA1c, and a diastolic blood pressure of 70. The three curves represent the survival curves for starting in one of the three transient stages. Figure 2 shows that the probabilities of remaining free of state 4 (Grade VI) retinopathy during a period of five years are 96%, 94%, and 86% starting from stage 1, 2 and 3 at time zero, respectively. These probabilities dramatically decrease during a period of 10 years to 77%, 75%, and 65%, respectively. These probabilities and Figure 2 also show that staying in stage 2 does not significantly increase the risk of progressing to diabetic retinopathy. However, stage 3 shows a significant reduction during the first five years of the probability of staying free of retinopathy and has similar reduction in the second five-year period when compared to the probabilities of stages 1 and 2.

8 Discussion

This paper has demonstrated that a multi-state Markov model is not only an innovative statistical tool for the analysis of longitudinal and event history data, but with the introduction of the PR and PMR models it is also a feasible regression technique.

The results of the multi-state model have confirmed much of what is known about the natural course and the factors affecting diabetic retinopathy. However, using the Markov model we have learned more about how the different factors affect the disease process over time.

As many cross-sectional studies have shown, duration of diabetes is the single most important factor associated with the rate of progression among the different stages of diabetic retinopathy. Some of the factors found to be significantly associated with eye complications were no longer significant when duration of diabetes was included in the model. In a multivariate model we found that cumulative mean HbA1c and diastolic blood pressure remained significant even after adjusting for the duration of the disease.

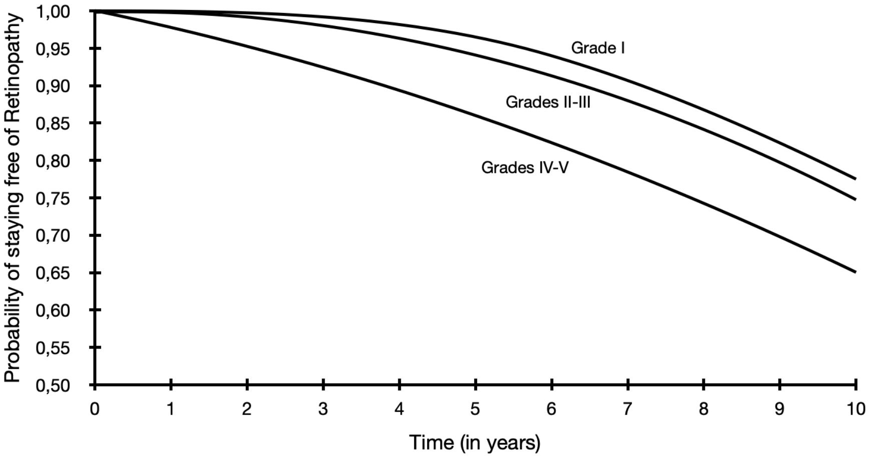


Fig. 2. Survival-type curves for the probability of staying free of grade VI retinopathy by eye grades

The regression analyses in the context of multi-state Markov models becomes a feasible statistical technique when the number of parameters associated with the different covariates are significantly reduced by introducing the PR and the PMR models. This not only prevents overfitting the data with redundant parameters, but also provides meaningful clinical information about the effects of different risk factors in the disease process. Almost all factors found to be significantly associated with diabetic retinopathy had their best representation in the PMR model. Duration of diabetes was not only the most important factor associated with changes in eye complications, but was the only variable for which the PR model was the best representation. Although age in years had its best representation in the full model, the likelihood ratio between the full model and the PR model was not significant at the 5% level (p=0.066).

Multi-state regression modelling should become an increasingly important and attractive statistical technique for the analysis of longitudinal data dealing with stages of chronic disease. However, this will only be possible when more computer programs for the models reviewed in this paper become more accessible and easy to use.

Bibliography

[1] Diabetic Retinopathy Study Research Group. A modification of the Airlie House classification of diabetic retinopathy: report 7. Invest Ophtalmol Vis Sci 1981, 21: 210-26.

[2] Klein BEK, Davis MD, Segal P, Long JA, Harris WA, Haug GA. Diabetic retinopathy: assessment of severity and progression. Ophthalmology 1984; 91:10-7.

[3] Sorbinil Retinopathy Trial Research Group. A Randomized Trial of Sorbinil, and Aldose Reductase Inhibitor, in Diabetic Retinopathy. Arch Ophthalmol. 1990; 108:1234-1244.

[4] Klein RK, Klein BEK, Moss SE, Davis MD, DeMets DL. Glycosylated Hemoglobin Predicts the Incidence and Progression of Diabetic Retinopathy. JAMA 1988; 260: 2864-2871.

[5] Teuscher A, Schnell H, Wilson PWF. Incidence of Diabetic Retinopathy and Relationship to Baseline Plasma Glucose and Blood Pressure. Diabetic Care 1988; 11: 246-251.

[6] Janka HU, Warram JH, Rand LI, Krolewski AS. Risk Factors for Progression of Background Retinopathy in Long-Standing IDDM. Diabetes 1989; 38: 460-64.

[7] The Diabetes Control and Complications Trial Research Group. The Effect of Intensive Treatment of Diabetes on the Development and Progression of Long- Term Complications in Insulin-Dependent Diabetes Mellitus. N Engl J Med 1993; 329: 977-86.

[8] Kay, R. (1986). A Markov Model for Analysing Cancer Markers and Disease States in Survival Studies. Biometrics 42, 855-865.

[9] Longini Jr.,I.M., Clark,W.S., Byers,R.H., Ward,J.W., Darrow, W.W., Lemp,G.F., and Hethcote, H.W. (1989). Statistical Analysis of the Stages of HIV Infection Using a Markov Model. Statistics in Medicine 8, 831-843. 15

[10] Garg, S.K., Marshall, G., Chase, H.P. et al. (1990). The Use of the Markov Processes in Describing the Natural Course of Diabetic Retinopathy. Arch. Ophthalmology, 108, 1245-1247.

[11] Marshall, G. Multi-State Markov Models in Survival Analysis, Ph.D. thesis (University of Colorado, Denver CO, 1990)

[12] Marshall, G. and Jones, R.H. Multi-State Survival Models. Technical Report 931, Department of Preventive Medicine and Biometrics, University of Colorado (1993).

[13] Marshall, G. and Jones, R.H. MARKOV: A Computer Program for Multi-State Markov Models with Covariables. Submmited to Computers and Biomedical Research.

[14] Cox, D.R. and Miller, H.D. The Theory of Stochastic Processes. Chapman & Hall, London, New York (1965).

[15] Dennis Jr., J.E. and Schnabel, R.B. Numerical Methods for Unconstrained Optimization and Nonlinear Equations. Prentice-Hall, Englewood Cliffs, New Jersey (1983).

Article 1.3

Ordinal Discrete Time Markov Transition Models

Guillermo Marshall and Wensheng Guo

Pontificia Universidad Católica de Chile and
University of Pennsylvania School of Medicine, Philadelphia, USA

Abstract. Markov transition models in equally spaced settings have been widely used in many applications. In equally spaced settings, the dependency from the previous observations are modeled explicitly by including the previous history as covariates. The estimation and inference are straightforward. However, in real practice such as data collected from a clinic, the observations are usually unequally spaced and missing data are common. The equally spaced Markov transition models break down in these general settings, because the transition probability not only depends on the previous stages, but also depends on the time interval between the observations. For rapid transition diseases, multiple transitions can happen if the interval between two observations are long enough. Another long existing problem for multi-stage Markov transition models is that the number of parameters expands exponentially with the increase of the number of states and covariates, which limits the application of such models to relatively simple cases. In this paper, we propose to model the transition rates as functions of covariates instead of the transition probabilities. We focus on the cases with ordinal outcomes, which can include absorbing stages such as dropout or death as special cases. We first propose general methodology and then propose a reduced common slope model that significantly reduces the number of parameters and increases the interpretability of the model. For computational convenience, we approximate the transition density matrix in the continuous time setting by the one-step transition probability matrix in a discrete time scale. This leads to an efficiently parametrized model and a computationally efficient estimation procedure. The model is illustrated using a four-state model for diabetic retinopathy in young subjects with insulin-dependent diabetes mellitus.

Keywords: Diabetic Retinopathy; Markov transition Model; Multi-State Model; Ordinal Response Models; Time-dependent Covariables. [1]

1 Introduction

Markov transition models have been widely used to model the progressions of chronic diseases such as cancer, AIDS, and diabetes. When patients are observed at equally spaced intervals, the dependency of the history can be modeled explicitly by including the previous observations as covariates. By doing so, one assumes that the exact transition time is observed. This has been the common practice of most of the Markov transition model.

[1] Partially published in Guo W and Marshall G. ORDMKV: A Computer Program Fitting Proportional Odds Model for Multi-State Markov Models. *Computers Methods and Programs in Biomedicine* 46: 257-263.

Unfortunately, in real practice patients visit a clinic at irregular time intervals and there are usually missing data even in the best controlled clinical trials. In such settings, modeling the transition probability by including the previous observations as covariate can be misleading, since (1) the longer the interval is, the more likely a transition may have occurred; (2) we seldom observe the exact transition time; (3) the patient may have been though several intermediate transitions between two consecutive observations. Therefore, we need to model the disease progression through the transition intensities instead of the transition probabilities.

In continuous time settings, the Markov transition model is completely specified by the transition intensity matrix. Kay (1) introduces a k-state Markov transition model with a special structure, in which a patient can progress to the adjacent stage or transfer to an absorbing stage. The exact transition times are not observed except in certain circumstances such as death. The model assumes that the underlying biological process can be characterized as a continuous-time Markov process and the covariates affect the disease progression through the transition densities.

Longini et al. (2) used this model to describe the distribution of the incubation period of HIV on patients with AIDS. Garg, Marshall, Chase et al. (6) used this model to describe the natural course of diabetic retinopathy. Marshall and Jones (3; 4) and Marshall, Guo and Jones (5) extended the k-state Markov transition model by proposing an exact likelihood function when the exact arrival time is observed and by introducing time-invariant and time-dependent covariables in the model. Marshall, Garg, Jackson et al. (7) used this regression model to identify various factors influencing the progression and regression of diabetic retinopathy in young subjects with type I diabetes mellitus. The limitations of these models are: (1) the model structure is specific; (2) the number of parameters increases exponentially as the number of stages and covariates increases; (3) it is computationally expensive to evaluate the likelihood in the continuous time setting; (4) it is difficult to interpret the parameters in terms of the overall progression, since each element in the transition intensity matrix has its own set of parameters.

In this paper we propose a general discrete time ordinal Markov transition model, because the stages of a disease are usually ordinal. In the discrete time setting, the progression of the disease is completely specified by the one-step transition probability matrix. When the time scale is chosen fine enough, it can be viewed as an approximation to the transition intensity matrix in the continuous time. The structure of the one-step transition probability matrix defines the possible transition paths of the disease and need to be specified according to the knowledge of the disease or the experimental design. When a path in the one-step transition is not allowed, its associated one-step transition probability is defined as a structural zero. A structural zero in the one-step transition probability matrix does not imply the same structural zero in the observed transition probability matrix. For example, we can restrict the disease to progress one stage at a time and therefore a direct progression from stage 1 to stage 3 is prohibited in the one-step transition. This translates into a structural zero probability for a transition from stage 1 to stage 3 in the one-step transition matrix. However, when the observation interval is longer than the underlying discrete time scale, it is possible to observed a transition from stage 1 to stage 3, simply because we do not observe the transitions from stage 1 to stage 2 and from stage 2 to stage 3. We then model each row of the one-step transition probability matrix as a proportional odds model, which reduces the number of parameters needed to characterize the disease progression significantly and increase the interpretability of these parameters. This is further extended to restrict different rows of the one-step transition probability matrix to share the same set of covariates and with same slops. This further reduces the number of parameters and makes the interpretation of the parameters even clearer. Further model reductions by putting constraints on the intercepts are discussed in section 2. These model assumptions can be easily checked by likelihood ratio tests because they are nested models.

A special case of particular interest is when there are absorbing stage(s) such as death or dropout. In many cases, death may be the worst stage. We are not only interested in the time to death, but also interested in the progression of the disease. This is particularly useful for quality of life data. In the cases with informative dropouts, a patient may drop out because the disease gets worse or because of remedy. In these cases, it is possible to include dropout as the worst or the best stage. An absorbing stage is defined by having a probability one to stay in its current stage and all the transition probabilities to other stages being zero in the one-step transition probability matrix. The difference between the dropout and death in terms our model is that we usually observe the exact transition time in the cases with death, but not with dropouts. They contribute differently to the likelihood, which will be further explained in section 4.

In the following of this paper, we describe the general model in section 2 and its estimation procedure in section 3. We describe some special cases with absorbing stage(s) in section 4. In section 5, we illustrate the model by an application to the diabetic retinopathy in young subjects with insulin-dependent diabetes mellitus. Discussion about limitations and future works are in section 6.

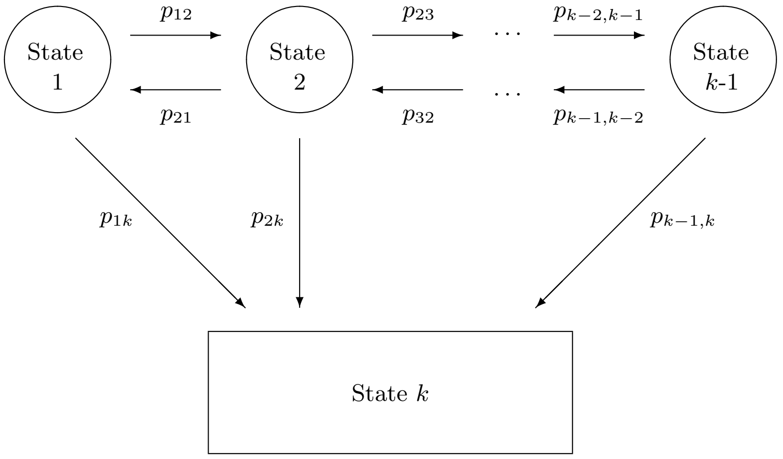
2 The Multi-State Markov Transition Model

For a biological process with k ordinal disease states, the progression or regression of the disease within a small interval (one-step transition) depends on the m previous stages as well as covariates. Such biological process can be modeled by a mth order Markov process. In the continuous time setting, the progression is completely specified by the transition intensities (8). Marshall and Jones (4) modeled these transition densities as functions of covariates. In the discrete time setting, the process is determined by the one-step transition matrix. We can then model the one-step transition probabilities as functions of covariates. For simplicity, we focus on first order Markov transition model in the current paper. Our method can be directly extended to incorporate high order Markov transition models by extending the dimension of the one-step transition matrix. When the one-step interval is chosen fine enough, the discrete time Markov process is a good approximation to the continuous time Markov process. However the finer the time grid, the more computationally expensive it is to estimate the parameters. In real practice, the grid is chosen large enough to ease the computational demand and small enough so that no more than one transition can happen within that interval.

Figure 1 shows Markov process with k stages. The paths represent possible one-step transition among different stages. Some of the paths can be forbidden due to biological or physical beliefs. An absorbing stage is a stage whose outgoing paths to other stages are all forbidden. The transition paths can be represented by the one-step transition probability matrix as

$$\mathbf{P}(\boldsymbol{z}) = \begin{bmatrix} p_{11}(\boldsymbol{z}) & p_{12}(\boldsymbol{z}) & \cdots & p_{1k}(\boldsymbol{z}) \\ p_{21}(\boldsymbol{z}) & p_{22}(\boldsymbol{z}) & \cdots & p_{2k}(\boldsymbol{z}) \\ \vdots & \vdots & \ddots & \vdots \\ p_{k1}(\boldsymbol{z}) & p_{k2}(\boldsymbol{z}) & \cdots & p_{kk}(\boldsymbol{z}) \end{bmatrix}, \qquad (1)$$

Fig. 1. The Multi-state Disease Process with associated one-step transition probabilities



where $\boldsymbol{z}$ is a set of covariates at the current time and each row of the transition probability matrix satisfies the basic property $p_{i1}(\boldsymbol{z}) + p_{i2}(\boldsymbol{z}) + \ldots + p_{ik}(\boldsymbol{z}) = 1$ $(i = 1, \ldots, k)$. The n-step transition probability for a subject with vector of covariates $\boldsymbol{z}$, denoted by $p_{ij}^{(n)}(\boldsymbol{z})$, represents the probability that the process is in the state j after n-steps given that the process started on state i at time t, that is

$$p_{ij}^{(n)}(\boldsymbol{z}) = Pr\{X(t+n) = j | X(t) = i, \boldsymbol{z}\}. \tag{2}$$

This can be evaluated by computing the nth power of the transition probability matrix $\mathbf{P}(\boldsymbol{z})$. The transition probability $p_{ij}^{(n)}(\boldsymbol{z})$ is the element (i, j) of the product matrix $\mathbf{P}^n(\boldsymbol{z})$. The evaluation of $\mathbf{P}^n(\boldsymbol{z})$ can be obtained as

$$\mathbf{P}^n(\boldsymbol{z}) = \mathbf{A}(\boldsymbol{z}) \begin{bmatrix} \rho_1(\boldsymbol{z})^n & 0 & \cdots & 0 \\ 0 & \rho_1(\boldsymbol{z})^n & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \rho_k(\boldsymbol{z})^n \end{bmatrix} \mathbf{A}^{-1}(\boldsymbol{z}), \tag{3}$$

where $\mathbf{A}(\boldsymbol{z})$ is the square matrix containing, in each column, the eigenvector associated with each of the eigenvalues, $\rho_i(\boldsymbol{z})$, of the transition probability matrix $\mathbf{P}(\boldsymbol{z})$.

From the formulation of the discrete-time Markov transition model, we can see that the model is characterized by the one-step transition probability matrix $P(\boldsymbol{z})$, which is indexed by the covariates $\boldsymbol{z}$. As mentioned in the introduction, the long-existing challenge for Markov transition models is to define an efficient parameterization so that the parameters can be stably estimated and have clear interpretations, because in the traditional parameterization there are k^2 elements in the one-step transition probability matrix $P(\boldsymbol{z})$ and each has its own set of parameters. In this section, we propose to model each row of $P(\boldsymbol{z})$ by a proportional odds model, and further reduce the model for the whole one-step transition probability matrix $P(\boldsymbol{z})$ to a “common slope model”, which enforces different rows to share same slope parameters. This approach reduces significantly the number of parameters needed to characterize the progression and provides clear interpretations for these parameters in terms of the effects of the covariates on the disease progression. Further model reduction by putting constraint on the intercept parameters will also be discussed.

We define the cumulative one-step transition probability as

$$\gamma_{ij}(\boldsymbol{z}) = \Pr\left\{X(t+1) \leq j | X(t) = i \mid z\right\}.$$

Using a proportional odds model, the cumulative one-step transition probabilities can be written as:

$$\gamma_{ij}(\boldsymbol{z}) = \frac{\exp\{\theta_{ij} - \beta_i' \boldsymbol{z}\}}{1 + \exp\{\theta_{ij} - \beta_i' \boldsymbol{z}\}}, \tag{4}$$

for $j = 1, \ldots, k-1$, and $\gamma_{ik}(\boldsymbol{z}) = 1$. The intercept parameters θ_{ij} are restricted to be ordered as $\theta_{i1} < \theta_{i2} < \cdots < \theta_{i,k-1}$, to preserve the ordinal structure of the model and the definition of a probability measure. The one-step transition probabilities can be calculated from the cumulative one-step transition probabilities as:

$$p_{ij}(\boldsymbol{z}) = \begin{cases} \gamma_{ij}(\boldsymbol{z}) & \text{if } \ j = 1 \\ \gamma_{ij}(\boldsymbol{z}) - \gamma_{i,j-1}(\boldsymbol{z}) & \text{if } \ j \neq 1 \end{cases}$$

Model (4) can be further reduced to a common slope model:

$$\gamma_{ij}(\boldsymbol{z}) = \frac{\exp\{\theta_{ij} - \beta' \boldsymbol{z}\}}{1 + \exp\{\theta_{ij} - \beta' \boldsymbol{z}\}}, \tag{5}$$

which implies the covariates have the same impact on the progression regardless of the patient's current stage. This assumption can be checked by a likelihood ratio test comparing models (4) and (5). A further reduction of the number of parameters can be introduced by modeling the intercepts θ_{ij} as $\theta_{ij} = \theta_i \alpha_i^{(i-j)}$ where $0 < \alpha_i < 1$, which can also be tested by a likelihood ratio test.

3 Parameter Estimation and the Likelihood Function

First we denote θ as the whole collection of all unknown parameters, which include the slope and intercept parameters defined above. In this section we describe how to estimate these parameters via maximum likelihood.

Suppose we observe ith $(i = 1, \ldots, m)$ at time t_{ij} with the stage s_{ij} and covariates $\boldsymbol{z}_{ij}$ $(j = 1, \ldots, L_i)$. The interval between two consecutive observations is calculated as $n_{ij} = t_{ij} - t_{i,j-1}$. In the absence of absorbing states, the log-likelihood can then be calculated as

$$l(\theta) = \sum_{i=1}^{m} \sum_{j=1}^{L_i} log\{p^{n_{ij}}_{s_{i,j-1}, s_{i,j}}(\boldsymbol{z}_{ij})\}, \tag{6}$$

where the transition probability $p^{n_{ij}}_{s_{i,j-1}, s_{i,j}}(\boldsymbol{z}_{ij})$ takes into account that multiple transitions can have happened within the interval $[s_{i,j-1}, s_{i,j}]$ and we do not observe the exact transition time.

In the presence of absorbing stages such as death, we may observe the exact transition time. The contribution to the likelihood needs to take into account this additional information explicitly. Suppose that we observe the exact transition time of *ith* subject to an absorbing stage s_{ij}. Using a similar formulation by Kay (1986), we can write the contribution to the likelihood function as

$$\bar{p}^{(n_{i,j})}_{s_{i,j-1},s_{ij}}(\boldsymbol{z}_{ij}) = \sum_{l \neq s_{ij}} p^{(n_{ij}-1)}_{s_{i,j-1},l}(\boldsymbol{z}_{ij}) p_{l,s_{ij}}(\boldsymbol{z}_{ij}). \tag{7}$$

Note that this contribution term explicitly forces the transition to the absorbing stage to happen within the last unit of time. We will further discuss some applications with absorbing stages in section 4.

The following outlines the steps in calculating the log-likelihood:

1) For a given value of θ, use model (4) or (5) to calculate cumulative one-step transition probabilities $\gamma_{ij}(\boldsymbol{z}_{ij}, \theta)$.

2) Calculate one-step transition probabilities $p_{ij}(\boldsymbol{z}_{ij}, \theta)$ from $\gamma_{ij}(\boldsymbol{z}_{ij}, \theta)$.

3) If we do not observe the exact transition time, calculate $\mathbf{P}^{n_{ij}}(\boldsymbol{z}_{ij}, \theta)$ using equation (3). The $(s_{i,j-1}, s_{ij})$ element of the $\mathbf{P}^{n_{ij}}(\boldsymbol{z}_{ij}, \theta)$ is the contribution to the likelihood from this transition.

4) If we observe the exact transition time, calculate $\mathbf{P}^{(n_{ij}-1)}(\boldsymbol{z}_{ij}, \theta)$ or using equation (3), and use equation (7) to calculate $\bar{p}_{s_{i,j-1},s_{ij}}$, which is the contribution to the likelihood from this transition.

5) Repeat step (2)-(4) for all transitions and sum over all the log transition probabilities using equation (6).

The log-likelihood can then be numerically maximized using a Quasi-Newton algorithm. Initial values can be critical in the process of finding the maximum likelihood estimates of the parameters. In the context of a continuous-time Markov model, a way to obtain the initial values was proposed by (1) to use the estimates of the parameters found by assuming that the observed times are the actual transition time of the process. We adopt the similar idea here. The details on how to obtain the initial values is given in the appendix.

4 Survival Curves

The survival function defined as the probability to stay in one of the transient state at time t,

$$S_i(t; z) = \Pr\{X(t) < k | X(0) = i \mid z\},$$

can be defined in the context of this discrete Markov model as $S_i(t; z) = 1 - p^{(t)}_{ik}(\boldsymbol{z})$. The overall survival function can be defined as

$$S(t; \boldsymbol{z}) = \sum_{i=1}^{k-1} \pi_i S_i(t; \boldsymbol{z}),$$

where π_i represents the probability to be at state i at any arbitrary time $t = 0$

5 Diabetic Retinopathy: An Example.

Two hundred seventy-seven subjects who had Type 1 diabetes for at least five years when initially seen in the Eye-Kidney Clinic at Barbara Davis Center for Childhood Diabetes were included in the analysis described by Marshall, Garg, Jackson et al. (7) using a continuous-time Markov model. In this study all subjects were seen at least twice with visits one or more years apart for a mean follow-up of almost 3 years. A grading of retinal findings for each visit was perform by a retinal specialist. The grades were based on the modified Airlie House classification in which grade I indicates no retinopathy; grade II indicates micro aneurysms only; grade III and IV indicate intermediate stages of background retinopathy, and grades V and VI indicate preproliferative and proliferative retinopathy, respectively. The worst eye grade was used for defining the state of the process in the Markov model. The Markov model was define regrouping the stages as follow:

$$\text{Grades I} \Leftrightarrow \text{Grades II-III} \Leftrightarrow \text{Grades IV-V} \Rightarrow \text{Grade VI.}$$

Different clinical factors related to the physiology and history of the disease were evaluated with respect to the influence on the onset, progression and regression of diabetic retinopathy. The most important factors related to changes in retinopathy were duration of diabetes, historical values of glycohemoglobin (HbA1) and distolic blood pressure.

A single-covariate, discrete-time Markov model was used to assess the individual effects of factors associated with diabetic retinopathy using a custom-designed computer program . The model with three regression coefficients (full model) and the model assuming common regression coefficients (restricted model) were fitted to each of the 10 factors considered in this study. Table 1 shows the fit of each model in terms of the -2 log likelihood function, the likelihood ratio test versus the null model, and the significance of the effect of each factor for both the full and restricted models. If a factor was found significantly associated with the progression or regression of diabetic retinopathy based on each of these models, the parsimonious representation of this factor was later used for multiple regression analysis.

The duration of diabetes, the age of the subject, and the mean HbA1 levels were factors most associated with transitions of diabetes retinopathy. Diastolic and systolic blood pressure, values of HbA1 at visit times, and smoking were also associated with changes in diabetic retinopathy. All factors except duration of diabetes and age are better represented by using the restrictive model that uses a common regression coefficient for all different transient states. More importantly, systolic blood pressure and smoking were found to be significantly associated with the disease process only when the restricted model was used, showing that the full model overfit the data and reduces the effect of these two factors by unnecessarily adding extra parameters.

The estimates of the regression coefficients of duration of diabetes are

$$(\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3) = (0.0539, 0.1152, 0.2165),$$

showing that the effect of duration increases over the course of the disease, in fact it doubles the effect from state 1 to state 2 and from state 2 to state 3.

Table 1. The -2 log-Likelihood function, the likelihood ratio test, and the P-value of single-covariable discrete-time Markov models for various factors associated with diabetic retinopathy using the full model (4), and the restrictive model $\boldsymbol{\beta}_i = \boldsymbol{\beta}$

.

	Full Model			Restrictive Model		
Factor	$-2l(\theta)$	$\chi^2(3)$	p-value	$-2l(\theta)$	$\chi^2(1)$	p-value
Duration of Diabetes	886.1	54.0	<0.001	892.5	47.6	<0.001
Age	912.4	27.8	<0.001	920.3	19.8	<0.001
Mean HbA$_{1c}$	916.6	23.6	<0.001	917.8	22.3	<0.001
Diastolic Blood Pressure	929.3	10.8	0.013	929.4	10.7	0.001
HbA$_{1c}$ at the Visit	929.4	10.7	0.013	931.0	9.1	0.003
Systolic Blood Pressure	933.3	6.8	0.079	933.8	6.3	0.012
Gender	934.6	5.6	0.136	938.5	1.6	0.203
Cholesterol	935.5	4.6	0.204	936.9	3.2	0.074
Smoking	936.0	4.1	0.251	936.2	3.9	0.048
Family Hx Hypertension	938.3	1.8	0.611	940.0	0.1	0.752

Table 2. Parameter estimates and standard errors for the best multiple regression model.

Factor	Parameter	Estimate	Standard Error
Constant	θ_{11}	7.643	0.7654
Constant	θ_{21}	1.172	0.7506
Constant	θ_{22}	9.760	0.0358
Constant	θ_{32}	3.775	1.1420
Constant	θ_{33}	13.24	0.0639
Duration of Diabetes	β_{11}	0.0667	0.0368
Duration of Diabetes	β_{12}	0.1286	0.0319
Duration of Diabetes	β_{13}	0.2193	0.0660
Mean HbA$_{1c}$	$\beta_{21} = \beta_{22} = \beta_{23}$	0.2364	0.0403
Diastolic Blood Pressure	$\beta_{31} = \beta_{32} = \beta_{33}$	0.1856	0.0058

Table 2 gives the estimates and the standard errors of the estimates for the parameters of the multiple regression model. Duration of diabetes remained the most important factor for explaining changes in diabetic retinopathy. As expected, cumulative mean values of HbA1 were the second most important clinical variable associated with transitions in retinopathy. Finally, diastolic blood pressure also remained in the model after adjusting for duration and HbA1 values, showing that it is an independent factor associated with the progression and regression of diabetic retinopathy.

6 Discussion

The use of the ordinal structure of the disease stages by introducing a proportional odds model in a discrete-time Markov chain successfully reduced the number of parameters involved in a multi-state disease model, and particularly compared to the representation of this model by a continuous-time Markov process. The estimation process and the inferences about the effect of factors affecting the progression and regression of the disease process gained in stability and consistency as shown with the example of diabetic retinopathy.

The results of our analysis of diabetic retinopathy shows that most of the factors related to the disease process can be represented by a simple restrictive model with only one parameter, and given the ordinal nature of this disease the loss of information is minimal. The other two factors were well-represented by the full model with three parameters which is still a reduction of the number of regression coefficients when compared to the continuous-time model.

By using the restrictive model in our example we found that two factors, smoking and systolic blood pressure, were significantly associated with changes in diabetic retinopathy. This type of finding is likely to occur in other applications and in those factors that while are not strongly associated with the disease process are still significant to clinicians.

This discrete model is a step forward to provide a feasible and meaningful methodology for the analysis of transitional data when exact times are not available, particularly to applications dealing with a large number of covariates and where the purpose is to find the best multiple regression model. More research will have to be done to compare the properties of this model with other link functions and to study the properties of the proportionality test using a continuous-time model as the reference.

Bibliography

[1] Kay. R. (1986). A Markov Model for Analyzing Cancer Markers and Disease States in Survival Studies. Biometrics 42: 855-865.

[2] Longini Jr., I. M., Clark, W. S., Byers, R. H., et al. (1989) Statistical Analysis of the Stages of HIV Infection Using a Markov Model. Statistics in Medicine 8: 831-843.

[3] Marshall G. and Jones R. H. (1993). Multi-State Survival Models. Technical Report. Department of Preventive Medicine and Biometrics, University of Colorado.

[4] Marshall G. and Jones R. H. (1995). Multi-State Markov Models and Diabetes Retinopathy. Statistics in Medicine. 14: 1975-1983

[5] Marshall G., Guo W., Jones R. H. (1995). Markov: A Computer Program for Multi-State Markov Models with Covariables. Computers Methods and Programs in Biomedicine. 47: 147-156

[6] Garg, S.K., Marshall, G., Chase, H. P., et al. (1990). The Use of the Markov Processes in Describing the Natural Course of Diabetic Retinopathy. Arch. Ophthalmology, 108: 1245-1247.

[7] Marshall, G Garg, S. K. Jackson, W. E., et al. (1993) Factors Influencing the Oncet and Progression of Diabetic Retinopathy in Subjects with Insulin-Dependent Diabetes Mellitus. Ophthalmology, 100: 1133-39

[8] Chiang, Ch. L. (1968) Introduction to Stochastic Processes in Biostatistics. Wiley, & Sons, Inc., New York, London,1968.

[9] Dennis Jr., J. E., Schnabel, R. B. (1983). Numerical Methods for Unconstrained Optimization and Nonlinear Equations. Prentice-Hall, Englewood Cliffs, New Jersey (1983).

[10] Cox, D. R. (1972). Regression Models and Life-Tables (with discussion). J.R. Statist. Soc.,B, 34: 187-220.

Article 1.4

MARKOV: A Computer Program for Multi-State Markov Models with Covariables

Guillermo Marshall, Wensheng Guo and Richard H. Jones

Pontificia Universidad Católica de Chile and
University of Colorado HSC, Denver, U.S.A.

Abstract. This paper discusses a computer program, called MARKOV, designed to fit a multi-state Markov model with covariables with a particular emphasis on the analysis of survival data. The Markov model consists of $k-1$ transient disease states and one absorbing state. The exact transition times are not observed except in situations such as death. Baseline transition intensities and regression coefficients are estimated via the method of maximum likelihood using a quasi-Newton optimization algorithm. The program's output includes the parameter estimates, the standard error of the estimates, the matrix of the correlation of the estimates and minus two times the log-likelihood function evaluated at the initial values and at the maximum likelihood estimates. Optionally, survival curves can be generated from each transient state, for one or more combination of covariables' values and simple tests about the parameters. The program is illustrated by using a four-state model to determine factors influencing diabetic retinopathy in young subjects with insulin-dependent diabetes mellitus.

Keywords: Multi-state models; Markov processes; survival analysis; quasi-Newton algorithm; time-dependent covariables; diabetic retinopathy. [1]

1 Introduction

Markov and semi-Markov models play an important role in the understanding and analysis of transition data from multi-state diseases such as cancer, AIDS, valvular heart disease and diabetes. Kay (1) introduces a general k-state Markov model in which the exact transition times are not observed except in certain circumstances such as death. Longini et al. (2) use this model to describe the distribution of the incubation period of HIV on patients with AIDS.

Marshall (3) and Marshall and Jones (4) extended this model in various directions. They introduced the exact likelihood function for continuous time processes, and they showed that this model is a generalization of parametric models in survival analysis. The model was also extended by introducing fixed and timed-dependent covariables similar to the proportional hazard model.

This paper describes a computer program designed to fit a general k-state Markov models with covariates and previously used by Garg, Marshall, Chase et al. (5) to describe the natural course of diabetic retinopathy, and by Marshall, Garg, Jackson et al. (6) to find various factors influencing

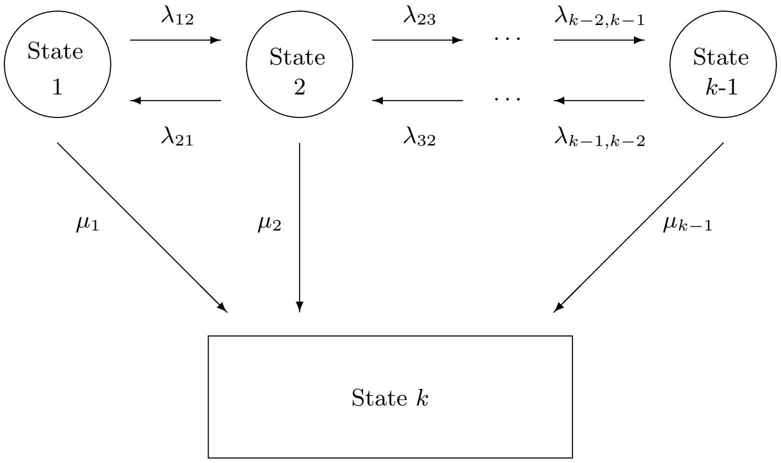
[1] Marshall G, Guo W, Jones RH. MARKOV: A Computer Program for Multi-State Markov Models with Covariables. Computers Methods and Programs in Biomedicine. 47: 147-156.

the progression and regression of diabetic retinopathy in young subjects with type I diabetes mellitus. The computer program was written in FORTRAN, and uses a set of optimization routines written by Schnabel, Koontz and Weiss (7).

2 The Model

A k-state Markov model includes k-1 transient disease states, $j = 1, 2, \ldots, k-1$, and one absorbing state k. The transient states are assumed to be ordered according to j and instantaneous transition represented by the intensities λ's and μ's, can take place from state j to the adjoining states $j-1$ or $j+1$. In addition to that, transition can take place from any state to the absorbing state k as is shown in Figure 1.

Fig. 1. A multi-state model with $k-1$ transient states and one absorbing state. The model has a total of $3k-5$ parameters, $2k-4$ λ's and $k-1$ μ's.



For the model in Figure 1 the transition intensity matrix **Q** can be written as

$$\mathbf{Q} = \begin{pmatrix} -(\mu_1 + \lambda_{12}) & \lambda_{12} & \cdots & 0 & \mu_1 \\ \lambda_{21} & -(\mu_2 + \lambda_{21} + \lambda_{23}) & \cdots & 0 & \mu_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & -(\mu_{k-1} + \lambda_{k-1,k-2}) & \mu_{k-1} \\ 0 & 0 & \cdots & 0 & 0 \end{pmatrix} \tag{1}$$

or the transition probability matrix P(t). The relation between the transition probability matrix P(t) and the transition intensity matrix Q is given by the Kolmogorov forward differential equations

$$\frac{d\mathbf{P}(t)}{dt} = \mathbf{P}(t)\mathbf{Q} \tag{2}$$

where the element (i, j) of the matrix $\mathbf{P(t)}$ represents the probability of a transition from the state i to the state j in a time interval t, denoted as $p_{ij}(t)$. We can express the solution to this system of differential equations as

$$\mathbf{P}(t) \;=\; \mathbf{A}\mathit{diag}\{\; e^{\rho_1 t}, e^{\rho_2 t}, \ldots, e^{\rho_k t} \;\}\mathbf{A}^{-1} \tag{3}$$

where $\mathbf{A}$ is the square matrix containing in column i the eigenvector associated with the eigenvalue ρ_i of the transition intensity matrix $\mathbf{Q}$.

Marshall (3) proved that given the form of the transition intensity matrix $\mathbf{Q}$, the eigenvalues always have real solutions and the matrix $\mathbf{Q}$ can be reduced to a tri-diagonal symmetric matrix by a similarity transformation. This property of $\mathbf{Q}$ allows the use of standard routines for finding the eigensystem of tri-diagonal symmetric matrices and significantly reduces the amount of computation time for the evaluation of the transition probability matrix, $\mathbf{P(t)}$, and, therefore, for the evaluation of the likelihood function.

The model can be extended by introducing covariables as a proportional factor over the baseline transition intensities as proposed in survival analysis by Cox (8). The regression model can be represented in terms of the element (i,j) of the transition intensity matrix $\mathbf{Q}$ as

$$q_{ij}(\mathbf{z}) = \lambda_{ij} e^{\beta'_{ij}\mathbf{z}} \tag{4}$$

where β_{ij} is the vector of regression coefficients associated with the vector of covariables $\mathbf{z}$ for the transition between the states i and j. The intensity matrix Q retains the property of the diagonal being minus the sum of the remain values in the same row. Marshall and Jones (4) introduced a new class of models by restricting the number of parameters allowing all progressive or all regressive transitions to have the same regression coefficients, that is,

$$q_{ij}\,(\mathbf{z}) = \begin{cases} \lambda_{ij} e^{\beta'_p \mathbf{z}} \; j = i + 1 \\ \lambda_{ij} e^{\beta'_r \mathbf{z}} \; j = i - 1 \end{cases} \tag{5}$$

An even more restrictive model can be introduced by setting the regression coefficients to $\beta_p = \beta_r = \beta$. Note that in both equations (4) and (5) $\lambda_{ik} = \mu_i$.

The model is related to survival analysis since the last state is an absorbing state. The functional relationship between the survival function and the transition probability matrix can be obtained from the equation $S_i(t|\mathbf{z}) = 1 - p_{ik}(t; \mathbf{z})$, where $S_i(t|\mathbf{z})$ is the survival function from the state i for a subject with covariables $\mathbf{z}$, and $p_{ik}(t; \mathbf{z})$ is the element (i, k) of the transition probability matrix $\mathbf{P}(t; \mathbf{z})$ evaluated using the intensities introduced in (4). Marshall (3) described further relationship of the Markov model with other important survival analysis functions such as: the density function, the hazard function, the mean lifetime, the median lifetime, and the residual mean lifetime.

The major distinction of this model compared to standard techniques is its ability to analyze unobserved transition times based on the observation of the process at arbitrary times. The typical information available from the patient are the states of the disease at each visit to the clinic and the exact time of death if death occurs. If i_1 and i_2 represent the observed states of the process at times t_1 and t_2, the contribution of this observation to the likelihood function is

$p_{i_1,i_2}(t_2 - t_1; \mathbf{z})$ for a regular transition, and

$$\sum_{j=1}^{k-1} p_{i_1,j}(t_2 - t_1; \mathbf{z}) q_{jk}(\mathbf{z})$$

when t_2 is the observed time of death. The total contribution to the likelihood function of an individual is the product of the contribution of each observation. The full likelihood is the product of the individual contributions. The likelihood function can be modified to handle time-dependent covariables. Constant covariables, $\mathbf{z}$, can be replaced by the value of the time-dependent covariables at the beginning of the interval, that is, by $\mathbf{z}(t_1)$.

Maximum likelihood estimates can be obtained by maximizing the likelihood function with respect to the parameters of the model, the λ's, μ's and β's, and asymptotic estimates of the standard error of the estimates can be obtained from the empirical information matrix. Quasi-Newton algorithms can be used to find the maximum likelihood estimates using only the likelihood function and finite differences to obtain numerical approximations of the derivatives, or by using explicit expression for the first derivatives. A complete discussion of these methods can be found in Dennis and Schnabel (9). These algorithms have been implemented by Schnabel, Koontz and Weiss [6] and in the IMSL FORTRAN library (10).

3 Aplications

The multi-state Markov model can be applied to follow-up medical studies dealing with the observation of transitions among multiple stages of chronic diseases. Particularly interesting are disease process with a low percentage of transitions to the final state, but with high number of transitions between intermediate states. In situations like this, standard survival analysis is very inefficient since most of the observation are right censored. This multi-state model can provide information about the progression and regression of the disease and the factors affecting this process.

The model has the flexibility to be applied to different applications and diseases processes. A basic model without covariables has been used by Longini et al. [2] to describe the HIV/AIDS disease process in which the states are HIV negative, HIV positive, AIDS related complex, AIDS and death. In this particular model, only transition in the direction of the progression of the disease are allowed.

A basic model also have been applied to diabetic retinopathy by Garg, Marshall, Chase et al. [5] with states representing increasing grades of complication and the absorbing state representing retinopathy. In this model bi-directional transitions are allowed, but the model does not allow direct transitions from the transient states to the absorbing state except from the preceding transient state $k - 1$. Marshall, Garg, Jackson et al. [6] use a model with covariables to determine factors influencing the process of diabetic retinopathy in the same population with insulin-dependent diabetes mellitus.

An application with a full set of transitions is a model for functional class on patients undergoing cardiac surgery. The functional class is measured on a scale from I to IV using the New York

Heart Association functional class to represent the transients states of the model and death, due to the cardiac disease, as the absorbing state. In this model bi-directional transition between the transients state, and direct transition from the transients states to the absorbing state are allowed.

4 The computer program

MARKOV was written in FORTRAN-77 and has approximately 2,500 lines of code. The program is also complemented by a few external optimization routines described by Schnabel, Koontz and Weiss (7). Using the new Microsoft 32-bit compiler, the current version of the program handles a data space of 40,000 observations , 10 states in the disease process and 10 different covariates. The program requires intensive computer resources mainly to evaluate the likelihood function. The contribution to the likelihood function of each two consecutive observations in each patient requires the evaluation of the transition probability matrix (3) for a given set of covariates z. The CPU time needed for MARKOV to fit a model varies according to the number of possible transitions in the model, the number of covariates, the number of observations, and the type of processor in the PC computer. The example presented in section 6 with five possible transitions, one covariate, 614 observations on 277 subjects, took a little more than 1 minute on a PC with an Intel 486 DX 66 Mhz processor.

5 Input/Output

The program's input is divided in two files; the first is the control file with default file extension MKV that contains the parameters and the specification of the model, including the names of the input/output files and parameters related to the definition of the Markov model. The second input file contains the actual data, including subject identification, original state of the transition, the final state of the transition, the transition time, and the values of the covariables. The name of this file is specified in the control file and the format of the data is assumed to be free.

MARKOV runs in batch mode by reading a control file containing a set of required and optional program statements. The required MARKOV statements are:

INFILE *filename;*
INPUT *variable-list;*
MODEL *varname* = *variable-list* [/*options*] ;
TIME *varname;*
ID *varname;*

The INFILE statement allows the user to specify the name of the file containing the data. Valid filenames include the use of drives, directories, name of the file, and file's extension. The INPUT statement describes the arrangement of the data in an input record and assigns these values to corresponding MARKOV variables. The order of the names of the variables should correspond with the order of the data assign to it. The data file should be in ASCII format and the columns separated by spaces.

The MODEL statement describes the model that the user wants to fit. The dependent variable specifies the variable listed in the INPUT statement that contains the state of the process at each given time. The independent variables specify the list of variables to be included in the model as covariates. The options of the model are:

nofit used to avoid that MARKOV fits a new model. This is particularly useful to compute survival curves after fitting a model, however the user has to provide the values of the parameters using the PARAMETERS statement,

exact used to control the type of contribution to the last state in the likelihood function. By default the likelihood function is calculated assuming that no exact transition time is observed from a transient state to the absorbing state. By adding this option the likelihood function will be calculated assuming that transition times to the last state are known, such as death.

likelihood prints in the screen the -2 log-likelihood function evaluated at the current values of the parameters in each iteration. This option allows the user to see the steps in the process of minimization.

history prints in the screen the values of the parameters, the gradient and minus two times the log-likelihood function at the current values in each iteration of the minimization process.

nocenter by default all covariates are centered about the mean. Generally by centering the covariates the user gets faster and more reliable results. This option specifies that the covariates should not be centered when entered in the model.

initial prints in the output file the initial values of the parameters.

correlation prints in the output the correlation of the estimates.

covariance prints in the output file the covariance of the estimates.

The ID statement specifies the variable in the INPUT statement that defined the subject identification variable. The TIME statement specifies the variable in the INPUT statement that defined the time of the process.

The optional statements of MARKOV are:

TITLE *text;*
MATRIX *entries;*
PARAMETERS *entries;*
CONSTRAINT *varname values* [,*varname values,* ...] ;
SURVIVAL *start-time end-time interval* [,*varname value* ...] ;
TEST *'label' varname contrast* [,*varname contrast,* ...];
OPTIONS *options* ;

The TITLE statement allows the user to specify a running title for the output file not longer than 80 characters. The text can contain any valid ASCII character except a semicolon, which is the reserved end-of-command character in MARKOV. The MATRIX statement is an optional statement that allows the user to specify a different model in terms of the transitions among states. By default, instantaneous transitions can take place from state j to the adjoining states $j-1$ and $j+1$ and the final state k.

The PARAMETERS statement allows the user to specify initial values of the parameters for the minimization algorithm. Although, MARKOV computes good initial values for the parameters, this statement allows the user to have control of those values. This is particularly important when local minimums are found. This statement can also be used to provide the values of the parameters in case a nofit option in the MODEL statement is used to evaluate the survival curve.

The CONSTRAINT statement allows the user to constrain the effect of the covariables in the different transitions. For example, suppose we have a four-state model with transitions (1 to 2, 2 to 1, 2 to 3, 3 to 2, and 3 to 4) and we want to force age to have the same regression coefficient for all progression transitions and for all regression transitions, then we use the statement

```
constraint age 1 2 1 2 1;
```

Suppose we force age to have the same regression coefficient for all progression and the same, but negative, coefficient for all regression transitions, then we change the CONSTRAINT statement as follows:

```
constraint age 1 -1 1 -1 1;
```

If you do not use the CONSTRAINT statement it is equivalent to use in this example as follow:

```
constraint age 1 2 3 4 5;
```

In the first case we have only two parameters associated to age. The first coefficient is shared by transitions 1 to 2, 2 to 3, and 3 to 4, and the second coefficient is shared by the transitions 2 to 1, and 3 to 2. In the second example we are forcing that the second coefficient in the first example be the negative value of the first coefficient. Finally in the third example we are specifying one coefficient for each transition associated with age for a total of five parameters. Given this is the default you do not need to specify the CONSTRAINT statement.

The SURVIVAL statement allows the user to compute survival curves from any of the transients states to the absorbing state k by providing values of the covariables included in the model. The start-time is the starting time period, end-time is the final time period, and interval is the length of the intervals in which the user wants to evaluate the survival-type curves. The varname and value allows the user to control the values of the covariables in which the curves will be evaluated. If no value of the covariables is specified, the mean value of the covariables is used.

The TEST statement allows the user to perform a Wald-type test about the regression parameters. Label is any text that identifies the test, the varname represents the name of the covariables for which the regression coefficients are being tested, contrast are the constants that form the contrast. For example, suppose we want to known if is worth it to fit a model using the CONSTRAINT statement forcing all the coefficient of age associated with progression of the disease to be equal and the same restriction for the regression transition; then we can run the following example:

```
test ‘Progression and Regression’ age 1 0 -1 0 0 , age 1 0 0 0 -1 , age 0 1 0 -1 0 ;
```

The OPTIONS statement allows the user to specify options of MARKOV. The same options listed in the model can be used in this statement. The options-list is one or more of the following options: nofit, exact, likelihood, history, nocenter, initial, correlation, and/or covariance.

6 Diabetic Retinopathy: An Example

Two hundred seventy-seven subjects who had Type 1 diabetes for at least five years when initially seen in the Eye-Kidney Clinic at Barbara Davis Center for Childhood Diabetes were included in the analysis described by Marshall, Garg, Jackson et al. [6].

In this study all subjects were seen at least twice with visits one or more years apart for a mean follow-up of almost 3 years. A grading of retinal findings for each visit was perform by a retinal specialist. The grades were based on the modified Airlie House classification in which grade I indicates no retinopathy; grade II indicates micro aneurysms only; grade III and IV indicate intermediate stages of background retinopathy, and grades V and VI indicate preproliferative and proliferative retinopathy, respectively. The worst eye grade was used for defining the state of the process in the Markov model. The Markov model was define regrouping the stages as follow:

$$\text{Grade I} \Leftrightarrow \text{Grades II-III} \Leftrightarrow \text{Grades IV-V} \Rightarrow \text{Grade VI}$$

Different clinical factors related to the physiology and history of the disease were evaluated with respect to the influence on the onset, progression and regression of diabetic retinopathy. One of the most important factors related to changes in retinopathy was glycohemoglobin (HbA1).

The control file needed to fit this model with HbA1 as a covariable is shown in Appendix A. A survival curve is requested for the first 10 years at one year intervals for a patient with constant HbA1 equal to 11.8 which is the average value observed in this group of subjects. In addition, two tests about the parameters are specified. The first tests for equal regression coefficients of the duration of diabetes of the progression transitions, and the second tests for equal regression coefficients for the regression transitions.

Appendix B shows the output of this sample run. The initial values of the parameters, the status of the optimization process, the maximum likelihood estimates, the correlation matrix, the survival curve and the statistical test are included.

7 Program Limitations

Some limitations apply to this version of MARKOV. The maximum number of states is limited to 10, however this number is adequate for most practical applications. The maximum number of covariables is restricted to be less than or equal than 10, and the total data space is restricted to 40,000 observations.

All these numbers can be changed very easily in the source code of MARKOV. These numbers are defined in MARKOV in the PARAMETER statement at the beginning of the main program and subroutines.

Bibliography

[1] R. Kay, A Markov Model for Analyzing Cancer Markers and Disease States in Survival Studies. Biometrics 42 (1986) 855-865.

[2] I.M. Longini Jr., W.S. Clark, R.H. Byers, J.W. Ward, W.W. Darrow, G.F. Lemp, and H.W. Hethcote, Statistical Analysis of the Stages of HIV Infection Using a Markov Model. Statistics in Medicine 8 (1989) 831-843.

[3] G. Marshall, Multi-State Markov Models in Survival Analysis, Ph.D. thesis (University of Colorado, Denver CO, 1990).

[4] G. Marshall, and R.H. Jones, Multi-State Markov Models and Diabetes Retinopathy. Statistics in Medicine. 14 (1995) (to appear)

[5] S.K. Garg, G. Marshall, H.P. Chase, et al., The Use of the Markov Processes in Describing the Natural Course of Diabetic Retinopathy. Arch. Ophthalmology, 108 (1990) 1245-1247.

[6] G. Marshall, S.K. Garg, W.E. Jackson, et al., Factors Influencing the Oncet and Progression of Diabetic Retinopathy in Subjects with Insulin-Dependent Diabetes Mellitus. Ophthalmology, 100 (1993) 1133-1139.

[7] Schnabel, Koontz and Weiss. A Modular System of Algorithms for Unconstrained Minimization. ACM Transactions on Mathematical Software, 11, (1985) 419-440.

[8] D.R. Cox, Regression Models and Life-Tables (with discussion). J.R. Statist. Soc.,B, 34, (1972) 187-220.

[9] J. E. Dennis Jr., and R.B. Schnabel Numerical Methods for Unconstrained Optimization and Nonlinear Equations. (1983) Prentice-Hall, Englewood Cliffs, New Jersey.

[10] IMSL MATH/LIBRARY User's Manual, Version 1.1 (1989) (IMSL Inc., 2500 CityWest Boulevard, Houston TX 77042.)

Appendix A

A sample run for the diabetic retinopathy model fitting glycohemoglobin (HbA1). The control file is:

```
TITLE Model for HbA1;
INFILE hba1.dat;
INPUT subject eyes time duration hba1 diastolic;
MODEL eyes =  hba1 / HISTORY initial correlation;
ID subject;
TIME time;
MATRIX 0 1 0 0,
       1 0 1 0,
       0 1 0 1,
       0 0 0 0;
SURVIVAL 0 120 12;
TEST 'Uniform progression' hba1 1 0 -1  0  0,
                           hba1 1 0  0  0 -1;
TEST 'Uniform regression'  hba1 0 1  0 -1  0;
```

Appendix B

Output of the model fitted using the control file in Appendix A:

```
MARKOV 4.0 - Multi-State Markov Models in Continuous-Time - April 1994
 Developed by Guillermo Marshall
 Department of Preventive Medicine and Biometrics
 University of Colorado - Health Sciences Center

 Model for HbA1
 Data file: alleye.dat
 Number of Records      :   891
 Number of Observations :   614
 Number of Subjects     :   277
 Number of States       :     4
 Number of Transitions  :     5
 Number of Covariables  :     1
 Number of Parameters   :    10

 Model: Eyes    = (Hba1    -  11.799)


Initial Values for the Parameters

  Variable Name    Transition         Values
  ----------------------------------------------
   Baseline         1  -->  2       .2212E-01
   Baseline         2  -->  1       .7837E-02
   Baseline         2  -->  3       .9645E-02
   Baseline         3  -->  2       .1948E-01
   Baseline         3  -->  4       .4175E-02
   Hba1             1  -->  2       .0000E+00
   Hba1             2  -->  1       .0000E+00
   Hba1             2  -->  3       .0000E+00
   Hba1             3  -->  2       .0000E+00
   Hba1             3  -->  4       .0000E+00
  ----------------------------------------------
  -2 Log-Likelihood :    971.7158


 Status of the Optimization Process

  Number of iterations is           :   20
  Number of function evaluations is :   26
  Number of gradient evaluations is :   21


 Maximum Likelihood Estimates of the Transition Intensity Matrix Q
```

```
  Variable Name    Transition       Estimates         Std Error
  ----------------------------------------------------------------
   Baseline         1  -->  2       .4261E-01         .5297E-02
   Baseline         2  -->  1       .1336E-01         .2426E-02
   Baseline         2  -->  3       .1455E-01         .2264E-02
   Baseline         3  -->  2       .2771E-01         .7581E-02
   Baseline         3  -->  4       .2619E-02         .1747E-02
   Hba1             1  -->  2       .2115E+00         .7174E-01
   Hba1             2  -->  1      -.2336E+00         .1316E+00
   Hba1             2  -->  3       .1083E+00         .9738E-01
   Hba1             3  -->  2      -.3069E-01         .1878E+00
   Hba1             3  -->  4       .5214E+00         .3239E+00
  ----------------------------------------------------------------
  -2 Log-Likelihood :    914.9807


 Correlation Matrix of the Estimates

  1.0000
   .3921 1.0000
  -.0190  .0280 1.0000
  -.0014  .0025  .4335 1.0000
  -.0005 -.0040 -.0238 -.0068 1.0000
   .0130  .0704  .0027 -.0011 -.0005 1.0000
   .0439  .2736  .0042  .0009 -.0005  .4315 1.0000
  -.0028 -.0035 -.2601 -.0533  .0068 -.0301  .0124 1.0000
  -.0016 -.0007 -.0458 -.1133  .0023 -.0021 -.0023  .4599 1.0000
   .0005  .0014  .0042 -.0018 -.7331 -.0003 -.0030 -.0414 -.0173 1.0000

Survival Functions

 Covariable Values:
 Hba1    =   .0000

 -------------------------------------
                           States
    Time             1       2       3
 -------------------------------------
     .00        1.0000  1.0000  1.0000
   12.00         .9997   .9978   .9734
   24.00         .9979   .9927   .9539
   36.00         .9944   .9862   .9386
   48.00         .9896   .9789   .9261
   60.00         .9836   .9712   .9153
   72.00         .9770   .9634   .9056
   84.00         .9698   .9555   .8967
   96.00         .9623   .9475   .8883
  108.00         .9547   .9395   .8802
  120.00         .9469   .9316   .8724
```

```
-------------------------------------

Wald Test:

Label: Uniform progression
Contrast:
  1  0 -1  0  0
  1  0  0  0 -1
 Chi-square:    1.7551   df:    2

Label: Uniform regression
Contrast:
  0  1  0 -1  0
 Chi-square:     .7810   df:    1
```

Article 1.5

Factors Influencing the Onset and Progression of Diabetic Retinopathy in Subjects with Insulin-dependent Diabetes Mellitus

Guillermo Marshall, Satish K. Garg, William E. Jackson, Douglas L. Holmes, H. Peter Chase

Pontificia Universidad Católica de Chile and
University of Colorado HSC, Denver, U.S.A.

Abstract.

Background: The etiology of diabetic retinopathy is poorly understood. In the current study, factors associated with the onset and the progression or regression of retinopathy are evaluated.
Methods: Two hundred seventy-seven subjects with insulin-dependent (type I) diabetes mellitus (IDDM) were evaluated longitudinally for retinal changes over a meanof 2.7 years. The multistate Markov model was used to analyze the influences of the duration of diabetes, a family history of hypertension, age, sex, cigarette smoking, systolic blood pressure, diastolic blood pressure, cholesterol levels, and longitudinal glycohemoglobin (GHb) values on the development and the progression or regression of retinopathy.
Results: Univariate analysis confirmed that four factors were significantly associated with the etiology and the progression or regression of diabetic retinopathy: age, duration of diabetes, mean longitudinal GHb levels (all at $P \leq 0.01$), and diastolic blood pressure ≤ 0.04). However, age was no longer significant when controlled by duration of diabetes. Cigarette smoking was only associated significantly with background retinopathy (stages 2 and 3). Systolic blood pressure, sex, a family history of hypertension, and colesterol levels were not significantly associated with retinopathy.
Conclusions: The onset of diabetic retinopathy is associated with the duration of diabetes, mean longitudinal GHb levels, smoking, and diastolic blood pressure. A longer duration of diabetes, higher GHb values, and higher diastolic blood pressure levels are associated with an increased risk of progression and a decreased chance of regression of diabetic retinopathy. Ophthalmology 1993;100:1133-1139 [1]

1 Introduction

The leading cause of new cases of blindness in persons 20 to 74 years of age in the United States is diabetic eye disease. (1) An estimated 97% of insulin-taking and 80% of noninsulin-taking persons with diabetes for 15 or more years have diabetic retinopathy. (2; 3) The natural course of diabetic retinopathy recently has been described. (4) Individual factors reported to be associated with diabetic retinopathy include duration of diabetes,(5) longitudinal glycohemoglobin (GHb) values, (5; 6) diastolic blood pressure, (7; 8) systolic blood pressure, age, sex, (9) colesterol

[1] Marshall G, Garg SK, Jackson WE, Holmes DL, Chase HP. Factors Influencing the Onset and Progression of Diabetic Retinopathy in Subjects with Insulin-Dependent Diabetes Mellitus. 1993. Ophthalmology, 100: 1133-39.

values,(10) limited joint mobility, (11) cigarette smoking, (12) and alterations in blood clotting factors. (13; 14) It is possible that some of these factors may act individually, whereas others may be a consequence of other alterations. It is thus important to determine the independent effect of individual factors on the development and the progression of diabetic retinopathy. In the current study, we report our analysis using multistate Markov models on 277 subjects evaluated longitudinally over a mean period of 2.7 years.

2 Subjects and Methods

Two hundred seventy-seven subjects who had type 1 diabetes for at least 5 years had a mean (± standard deviation) age of 17.9 ± 3.2 years (range, 14-29 years) when initially seen in our Eye-Kidney Clinic (Table 1). The Eye-Kidney Clinic' is open to all patients who are 14 years of age or older and who have had type 1 diabetes for at least 3 years. There are no other criteria for attending this clinic, and more than 95% of eligible patients (followed in our general diabetes clinic) attend the Eye-Kidney Clinic. All subjects signed a consent form approved by the University of Colorado Health Sciences Center Human Subjects Committee.

There were 138 males and 139 females (Table 2), and the mean duration of insulin-dependent (type I) diabetes mellitus (IDDM) was 9.7 ± 3.9 years (range, 3-28) years. The mean or the distribution of other demographic variables also are reported in Tables 1 and 2. In the current study, all subjects were seen at least twice with visits 1 or more years apart, for a mean follow-up of 2.7 years (range, 1-9 years). One hundred five subjects (38%) attended the Eye-Kidney Clinic twice, 84 subjects (30%) attended three times, 45 subjects (16%) attended four times. 24 subjects (9%) attended five times, and 19 subjects (7%) attended six or more times. A total of 882 patient visits occurred, allowing 614 transition observations of change or no change in grade of retinopathy. Among the patients with only two visits, 52% were within 1 year and 48% were in a period of more than 1 year. The average time between visits among these patients was slightly more than 14 months. For patients attending the clinic more than twice, 40% of the visits were within 1 year, and 60% were separated by more than 1 year. The average time between visits for these patients was slightly less than 15 months.

Table 1. Mean and Standard Deviation of Some Demographic and Clinical Variables for 277 Patients with Insulin-dependent Diabetes Mellitus at the Beginning of the Study

Variable	Mean ± SD
Duration of diabetes (yrs)	9.7 ± 3.9
Age (yrs)	17.9 ± 3.2
Mean HbA1 (%)	11.8 ± 1.7
HbA1 at the visit (%)	11.4 ± 2.1
Diastolic blood pressure (mmHg)	68.3 ± 11.2
Systolic blood pressure (mmHg)	115.1 ± 13.4
Cholesterol (mM)	4.8 ± 1.1
SD = Standard Deviation	HbA1 = hemoglobin A1

The ophthalmologic examinations included pupil dilation, direct ophthalmoscopy by two examiners (1 ophthalmologist and 1 pediatrician), a slit-lamp examination, seven standard field color retinal photographs of both eyes, and fluorescein angiography. The initial retinal grading was performed by the ophthalmologist during the direct ophthalmologic/slit-lamp examination. The final grading for each eye at each clinic visit was accomplished using photographs of the seven standard fields and (for initial evaluations or when marked changes were present) fluorescein angiograms. The grading of retinal findings for each visit was performed by the retinal specialist in a masked fashion without knowledge of previous retinal classification or other clinical or laboratory parameters. The retinal specialist graded retinal findings using a modified Airlie House classification, (15; 16) in which grade I indicates no retinopathy; grade Il indicates microaneurysms only; grades Ill and IV indicate intermediate stages of background retinopathy; and grades V and VI indicate preproliferative and proliferative retinopathy, respectively. The worse eye grade for each visit was used to classify subjects and for statistical analysis purposes. Progression (worsening) and regression (improvement) in eye grades were defined as corresponding changes between any of the four stages of retinopathy created for the Markov model (as defined below) and observed at the follow-up visits. The four stages of the Markov model were defined as grade I as stage 1; grades Il to III as stage 2; grades IV to V as stage 3; and grade VI as the absorbing stage 4. Absorbing state is defined as the stage from which there is no regression of retinopathy. With this definition, we reduce significantly nonexistent transitions observed due to potential misclassification.

Serial GHb values were measured using ion exchange resin (Isolab Fast Hemoglobin Test System, Akron, OH). Normal values remained the same(6.3%-8.2%) throughout the study period. For the purposes of this study, all individual GHb values for each patient were used in statistical analyses, although only the value at the time of the visit and the mean (± standard deviation) value are shown in Table 1.

Table 2. Distribution of Some Demographic and Clinical Variables for 277 Patients with Insulin-dependent Diabetes Mellitus at the Beginning of the Study

Factors	Frequency (%)
Sex	
M	138 (50)
F	139 (50)
Smoking	
Never	187 (68)
Past	31 (11)
Now	59 (21)
Family History of Hypertension	
No	184 (66)
Yes	78 (28)
Unknown	15 (6)

Blood pressure was measured using a conventional mercury sphygmomanometer and the appropriate sized cuff at each visit after the subjects rested in a sitting position for 5 minutes. All blood pressure recordings were used in the Markov model (although only the mean [± standard deviation] values are given in Table1). Serum colesterol levels were measured yearly by enzymatic assay (peroxidase-phenol-4-amino-phenazone, Boehringer Mannheim Diagnostics, Mannheim, Germany) using the Boehringer Mannheim/Hitachi 717 analyzer (Indianapolis, IN). The mean (± standard deviation) of all previous recorded values is shown in Table 1. Cigarette smoking was defined as described previously (17) however, a dichotomous indicator of ever having smoked was used for the Markov modeling. The family history for hypertension was considered positive if any first-degree relative had received medication for the treatment of hypertension (≥ 141/90) before 50 years of age.

3 Statistical Methods

The SAS program package (19) was used to calculate descriptive statistics, and one-way analysis of variance and the chi-square test were used to study the association of all factors with the initial stages of diabetic retinopathy. The multistate Markov model, first introduced for a basic model (19) and later extended to include covariates, (20) was used to analyze univariate and multivariate effects of clinical factors on the transition of the various stages of diabetic retinopathy. This transition can take place in both directions for all grades of diabetic retinopathy except grade VI, which is defined as the absorbing state (stage 4) from which there is no regression of diabetic retinopathy. The stages of the Markov model, the path of possible transitions, and the transition rates associated with each event are illustrated in Figure 1. In contrast to other more conventional statistical models, such as logistic regression, this model can analyze the data longitudinally and assess the effects of clinical factors over the course of the disease process. Another major distinction of this multistate Markov model, with respect to other longitudinal analyses, is its ability to analyze unobserved transition times based on the observation of the process at arbitrary times. The typical information collected at each visit from the patient is the grade of diabetic retinopathy and other disease re- lated measurements.

The multistate model for diabetic retinopathy shown in Figure 1 can be extended by introducing covariates in the transition rates. Each transition rate can be modeled similarly to the Cox regression model in survival analysis, (21) with the exception that the hazard function is considered here as the transition rate and is assumed to be constant over time. The parameters of the model for the five transitions are estimated simultaneously via the method of maximum likelihood, using a custom-designed computer program, as previously described. (20) The model also allows the calculation of a transition probability matrix that can be used to predict the probability of transition from and to any two stages, and particularly the probability of progression to the absorbing stage (grade VI) from the earlier stages of diabetic retinopathy. The transition probability matrix is related to the transition rates by a system of differential equations, similar to the funcional relationship between the survival function and the hazard function in classic survival analysis. A closed-form solution for the transition probability matrix is available when the transition rates are constant over time. These transition probabilities can be calculated for any time and for any combination of risk factors. Survival-type curves can be obtained for the probability of not reaching the absorbing state. (20)

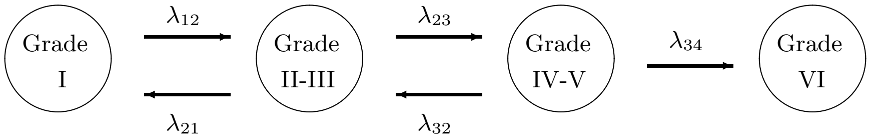


Fig. 1. The multistate Markov models for four stages of diabetic retinopathy defined on the basis of eye grades according to the Airlie House classification. The parameters (λs) represent the intensity of transitions of progression or regression.

4 Results

At the beginning of the study, the patients were distributed among the four stages of diabetic retinopathy as follows: stage 1 (grade I), 42%; stage 2 (grades II and III), 53%; stage 3 (grades IV and V), 5%; and stage 4 (grade VI), by definition of the analysis started with no subiects. The distribution of subjects at the end of the study period was as follows: stage 1, 26%; stage 2, 53%; stage 3, 19%; and stage 4, only 2%.

Among the 614 transition observations. 198 (32%) were observed from stage 1, 336 (55%) occurred from stage 2, and 80 (13%) from stage 3. Among those transitions from stage 1, 59%, 37%, 4%, and 0% were designated to stages 1, 2, 3, and 4, respectively. Among the transitions from stage 2, 12%, 73%, 14%, and 1% were designated to stages 1, 2, 3, and 4, respectively. Finally; 0%, 17%, 79%, and 4% of the transitions from stage 3 were designated to stages 1, 2, 3, and 4, respectively.

Subjects with higher grades of retinopathy at the beginning of the study were older and had a longer mean duration of diabetes than did subjects without retinopathy (Tables 3 and 4). Mean GHb, a history of smoking, systolic blood pressure, and diastolic blood pressure also were significantly different among subjects with the first three stages of diabetic retinopathy. Mean colesterol levels, the GHb at the time of the visit, a family history of hypertension, and sex were not significantly different among the various grades of diabetic retinopathy.

Table 3. Mean (± standard error of the mean) of Some Demographic and Clinical Variables by Stages of Diabetic Retinopathy at the Beginning of the Study

	Stages of Diabetic Retinopathy		
	1	2	3
Variables	(n=115)	(n=148)	(n=14)
Duration of diabetes (yrs)*	8.0 ± 0.3	10.6 ± 0.3	13.5 ± 1.1
Age (yrs)*	16.9 ± 0.3	18.2 ± 0.2	22.0 ± 1.3
Mean HbA1 (%)*	11.3 ± 0.2	12.1 ± 0.1	12.4 ± 0.5
HbA1 at the visit (%)	11.3 ± 0.2	11.4 ± 0.2	11.6 ± 0.6
Diastolic blood pressure (mmHg)*	65.8 ± 0.9	69.1 ± 0.9	80.6 ± 3.4
Systolic blood pressure (mmHg)*	113.6 ± 1.3	115.5 ± 1.0	123.7 ± 5.6
Cholesterol (mM)	4.9 ± 0.1	4.7 ± 0.1	5.5 ± 0.5

HbA1 = hemoglobin A1
* Significant differences ($P \leq 0.05$) among the three stages using an analysis of variance F test.

Univariate Markov models were used to assess the individual effect of factors associated with diabetic retinopathy using a custom-designed computer program (Table 5). Three different models were fit for each factor considered in this study. The first model asumes that each factor has a different effect on each one of the five posible transitions between the stages of diabetic retinopathy as shown in Figure 1; the second model asumes that each factor has the same effect in all progressive transitions and the same effect in all regressive transitions; and, finally, the third model assumes that the effect in the progressive transitions is inversely related to the regressive transitions. The importance of finding the parsimonious model among these three is crucial to best represent the association between a single factor and the disease process. In doing this, we can reduce from five to two (and 1 in other cases) the number of regression coefficients associated with each risk factor, and therefore, better assess and understand the effect on the transition times. If the factor was found to have a significant effect on diabetic retinopathy based on univariate analysis, the best representation among these three models was used for future multivariate analysis. The selection of the best model was performed using likelihood ratio test.

Table 4. Distribution of Some Demographic and Clinical Variables by Stages of Diabetic Retinopathy at the Beginning of the Study

	Stages of Diabetic Retinopathy*		
	1	2	3
Factor/Categories	(n=115)	(n=148)	(n=14)
Sex			
M	53 (46%)	62 (54%)	77 (52%)
F	71 (48%)	8 (57%)	6 (43%)
Smoking†			
Never	91 (79%)	89 (60%)	7 (50%)
Past	6 (5%)	24 (16%)	1 (7%)
Now	18 (16%)	35 (24%)	6 (43%)
Family History of Hypertension			
No	83 (72%)	93 (64%)	8 (57%)
Yes	30 (26%)	44 (30%)	4 (29%)
Unknown	2 (2%)	9 (6%)	2 (14%)

*Each cell contains the frequency and the percentage in parenthesis.
†Significant differences ($P \leq 0.05$) among the three stages using Chi-square test.

The age of the subject, duration of diabetes, and mean GHb levels were significantly associated ($P \leq 0.01$) with the five posible transitions (progression of stages 1 to 2, 2 to 3, or 3 to 4, or regression of stages 3 to 2 or 2 to 1) of diabetic retinopathy. Diastolic blood pressure also was significant at a level of $P \leq 0.04$ (Table 6). All individual significance levels were calculated using the likelihood ratio chi-square test comparing the univariate model with the basic model without covariables. All other factors, including sex, mean colesterol levels, family history of hypertension, systolic blood pressure, and a history of smoking were not significantly associated with changes in diabetic retinopathy stages.

The relative risk of factors on the progression (stages 1 to 2, 2 to 3, and 3 to 4) and regression (stages 3 to 2 and 2 to 1) of diabetic retinopathy was calculated with a 95% confidence interval using univariate and multi variate analyses (Table7). Duration of diabetes was important for regression of diabetic retinopathy, because the addition of I year decreased the chance of regression by 20%, whereas the progression risk increased only by 5% (Table 7). Also, if the duration of diabetes was longer, the chance of diabetic retinopathy progression was higher. The opposite also was true; that is, there was a greater chance of regression of diabetic retinopathy for someone having a shorter duration of diabetes.

Table 5. Univariate Analysis of Various Factors Associated with Diabetic Retinopathy Using the Markov Model

Factor	Chi-square*	P-value	Association†
Duration of diabetes (yrs)*	58.2	< 0.01	Positive
Age (yrs)*	33.5	< 0.01	Positive
Mean HbA1	27.2	< 0.01	Positive
Diastolic blood pressure (mmHg)	12.0	0.04	Positive
HbA1 at the visit	10.7	0.06	Positive
Sex	9.8	0.08	Inconclusive
Smoking	9.4	0.09	Positive
Systolic blood pressure (mmHg)	6.7	0.24	Positive
Cholesterol (mM)	5.0	0.42	Positive
Family history of hypertension	4.5	0.48	Inconclusive

HbA1 = hemoglobin A1
* Likelihood ratio test using the Markov model
†The association between the different factors and diabetic retinopathy.

Table 6. Individual Effects and Significance Levels in Multivariate Models of Selected Factors Found To Be Associated Significantly with Diabetic Retinopathy in the Univariate Analysis

Factors	Controlled by	Chi-Square	P-value
Duration of diabetes	Mean HbA1	61.7	< 0.01
Mean HbA1	Duration of diabetes	30.7	< 0.01
Diastolic blood pressure	Duration and Mean HbA1	10.7	0.04

HbA1 = hemoglobin A1
* Likelihood ratio test using the Markov model

5 Discussion

The multistate Markov model, using the continuous-time Markov process as described earlier, (19) was used to assess the longitudinal effects of various factors on the changes of diabetic retinopathy. Application of this model to the retinal data collected on these subjects with type I diabetes showed that some factors (duration of diabetes and longitudinal GHb levels) are important in the initiation of retinopathy. Cigarette smoking and diastolic blood pressure (in addition to duration of diabetes and GHb values) also are important in the progression of diabetic retinopathy. This is the first report, using these new statistical techniques, that describes factors that may influence the onset of diabetic retinopathy as well as those that may be important in the progression or regression of diabetic retinopathy. Factors not associated with either the initiation or the progression of retinopathy included sex, a family history of hypertension, and mean cholesterol levels. Mean cholesterol levels and sex have previous been reported to be associated with diabetic retinopathy.(9; 10)

The difference in the results when a single GHb value from the date of the eye evaluation was used for analysis ($P \leq 0.05$) versus when all longitudinal GHb values were considered ($P \leq 0.01$) is not surprising. The eye damage occurs over many years and is not an acute phenomenon observed just on the day of the clinic visit. This most likely explains the composite effect of por metabolic control as described previously. (5; 6; 22)

Table 7. Relative Risk from Various Factors on the Progression and Regression of Diabetic Retinopathy Using Univariate and Multivariate Analyses

	Univariate		Univariate		Unit of
Factors	Progression	Regression	Progression	Regression	Change
Duration of diabetes	1.054	0.800	1.075	0.812	1 yr
	(1.00, 1.11)*	(0.73, 0.88)	(1.02, 1.14)	(0.74, 0.89)	
Mean HbA1	1.996	0.836	1.237	0.808	1%
	(1.11, 1.29)	(0.78, 0.90)	(1.15, 1.33)	(0.75, 0.87)	
Diastolic blood pressure	1.201	0.832	1.195	0.837	10 mmHg
	(1.08, 1.34)	(0.75, 0.93)	(1.07, 1.33)	(0.75, 0.93)	
Smoking	1.871†	0.604	1.416	0.706	Yes/No
	(1.12, 2.46)	(0.41, 0.89)	(0.97, 2.06)	(0.49, 1.03)	

HbA1 = hemoglobin A1

* 95% confidence interval.

† Progression from stages 2-3 and regression from stages 3-2 only.

In the current report, we confirm our initial observation4 that diabetic retinopathy can improve as well as worsen. Nine percent of patients with diabetic retinopathy showed improvement, whereas 22% showed worsening. This study documents a relationship between longitudinal GHb levels (and age and duration of diabetes) with progression or regression of diabetic retinopathy. Although the Diabetes Control and Complications Trial will allow evaluation of factors influencing retinopathy, the GHb values of the subjects in the study are remaining quite low and quite consistent.23 Evaluation of the effects of fluctuating GHb levels during the Diabetes Control and Complications Trial period of observation will be difficult.

The duration of diabetes remained a significant factor when the effects of all other factors were removed. This most likely reflects the influence of currently unrecognized factors. A posible factor, as evidenced by retinal improvement after ablation of the pituitary gland,24 relates to levels of growth hormone (or other pituitary-related hormones or growth factors). Another likely influence, not evaluated in the current study, is that of blood clotting. ’° Future studies should provide further evaluation of these factors and should search for new previously unrecognized influences on the course of diabetic retinopathy. In the meantime, physicians should counsel patients with IDDM against smoking, and should vigorously at- tempt to control diastolic blood pressure and blood glucose levels.

Bibliography

[1] Klein R, Klein BEK. Vision disorders in diabetes. In: Na- tional Diabetes Data Group. Diabetes in America: Diabetes Data Compiled 1984. [Bethesda, MD]: US Dept. of Health and Human Services, 1985; chap. XIII.

[2] Klein R, Klein BEK, Moss SE, et al. The Wisconsin Epi- demiologic Study of Diabetic Retinopathy. Il. Prevalence and risk of diabetic retinopathy when age at diagnosis is less than 30 years. Arch Ophthalmol 1984;102:520-6.

[3] Klein R, Klein BEK, Moss SE, et al. The Wisconsin Epi- demiologic Study of Diabetic Retinopathy. III. Prevalence and risk of diabetic retinopathy when age at diagnosis is 30 or more years. Arch Ophthalmol 1984;102:527-32.

[4] Garg SK, Marshall G, Chase HP, et al. The use of the Mar- kov process in describing the natural course of diabetic ret- inopathy. Arch Ophthalmol 1990;108:1245-7.

[5] Chase HP, Jackson WE, Hoops SL, et al. Glucose control and the renal and retinal complications of insulin-dependent diabetes. JAMA 1989;261:1155-60.

[6] Klein R, Klein BEK, Moss SE, et al. Glycosylated hemo- globin predicts the incidence and progression of diabetic retinopathy. JAMA 1988;260:2864-71.

[7] Chase HP, Garg SK, Jackson WE, et al. Blood pressure and retinopathy in type I diabetes. Ophthalmology 1990;97:155-9.

[8] Harrold BP. Diabetic retinopathy and hypertension. Br J Ophthalmol1971;55:225-32.

[9] Bodansky HJ, Cudworth AG, Drury PL, Kohner EM. Risk factors associated with severe proliferative retinopathy in insulin-dependent diabetes mellitus. Diabetes Care 1982;5: 97-100.

[10] Miccoli R, Odello G, Giampietro O, et al. Circulating lipid levels and severity of diabetic retinopathy in type I diabetes mellitus. Ophthalmic Res 1987;19:52-6.

[11] Garg SK, Chase HP, Marshall G, et al. Limited joint mobility in subjects with insulin dependent diabetes mellitus: rela- tionship with eye and kidney complications. Arch Dis Child 1992;67:96-9.

[12] Klein R, Klein BEK, Davis MD. Is cigarette smoking as- sociated with diabetic retinopathy? Am J Epidemiol 1983;118:228-38.

[13] McMillan DE. Plasma protein changes, blood viscosity, and diabetic microangiopathy. Diabetes 1976;25(Suppl 2):858- 64.

[14] Miller JA, Gravallese E, Bunn HF. Nonenzymatic glycos- ylation of erythrocyte membrane proteins. Relevance to di- abetes. J Clin Invest 1980;65:896-901.

[15] Diabetic Retinopathy Study Research Group. Report 7. A modification of the Airlie House classification of dia- betic retinopathy. Invest Ophthalmol Vis Sci 1981;21: 210-26.

[16] Klein BEK, Davis MD, Segal P, et al. Diabetic retinopathy: assessment of severity and progression. Ophthalmology 1984;91:10-7.

[17] Chase HP, Garg SK, Marshall G, et al. Cigarette smoking increases the risk of albuminuria among subjects with type I diabetes. JAMA 1991; 265: 614-7.

[18] SAS User's Guide. Statistics. Version 5 ed. Cary, NC: SAS Institute,1985.

[19] Kay R. A Markov model for analysing cancer markers and disease states in survival studies. Biometrics 1986;42:855. 65.

[20] Marshall G. Multistate Markov Models in Survival Analysis (dissertation). Denver: Univ Colorado Health Sciences Center, 1990.

[21] Cox DR, Oakes D. Analysis of Survival Data. London: Chapman and Hall, 1984.

[22] Doft BH, Kingsley LA, Orchard TJ, et al. The association between long-term diabetic control and early retinopathy. Ophthalmology1984;91:763-9.

[23] DCCT Research Group, Diabetes Control and Complications Trial (DCCT). Update. Diabetes Care 1990;13:427-33.

[24] Foster DW. Diabetes mellitus. In: Wilson JD, Braunwald E, Isselbacher KJ, et al, eds. Harrison's Principles of Internal Medicine, 12th ed. New York: McGraw-Hill, 1991; 1739-59.

Chapter 2

Classification with Unbalanced Longitudinal Data

Article 2.1

Linear discriminant models for unbalanced longitudinal data

Guillermo Marshall and Anna E. Barón

Pontificia Universidad Católica de Chile and
University of Colorado HSC, Denver, U.S.A.

Abstract. This paper discusses statistical methods for the classification of observations into one of two or more groups based on longitudinal observations. Measurements on subjects in longitudinal medical studies are often collected at different times and on a different number of occasions. Classical multivariate methods for linear discriminant analysis are difficult to apply to repeated measurements due to the highly unbalanced structure observed in these data. Linear models for the analysis of longitudinal data proposed by Laird and Ware and non-linear models proposed by Lindstrom and Bates can be used to estimate population parameters for a discriminant model that classifies individuals into distinct predefined groups or populations. An example is presented using data from a study in 150 pregnant women in Santiago, Chile, in order to predict normal versus abnormal pregnancy outcomes.

Keywords: Discriminant Analysis; Multivariate Longitudinal Data; Linear and Nonlinear random effects models [1]

1 Introduction

Classical linear discriminant analysis has been used to classify subjects into one of g groups or populations using multivariate observations. Commonly, these vectors of multivariate observations are obtained from cross-sectional studies and represent different subject characteristics such as age, gender or other relevant factors. In general, a common and unrestricted covariance matrix is assumed in the g different groups.

Modifcations of this method have also been used to classify subjects when the vector of multivariate observations represents repeated measures collected in a longitudinal study. Azen and Afifi [1] introduced a two-stage model in which a discriminant function is obtained at each time point and the coeffcients entered into a linear regression of the function versus time to obtain a slope and intercept. The slopes and intercepts are used as input data to a final linear discriminant function. This method is limited by the fact that multiple observations per subject are required in order to allocate a subject to one of g groups at any point in time.

Albert et al. [2; 3] proposed a discriminant model for longitudinal data based on response curves. If only discrete measurements are collected in the study, the response curve can be estimated

[1] Marshall G and Barón AE. (2000) Linear Discriminant Analysis for Unbalanced Longitudinal Data. *Statistics in Medicine*, 19, 1969-1981

by linear interpolation or another type of curve fitting. A risk index based on the distance of an individual from a population response curve is used to classify observations. The major disadvantage to this approach is the same as that of the two-stage method described above, that is, the need for several observations per subject in order to estimate the underlying response curve and to classify subjects into groups.

Albert and Kshirsagar [4] proposed an exploratory method based on a growth curve structure embedded in a canonical variate analysis to achieve dimension reduction in a discriminant analysis framework. The authors suggested use of this approach for classification but did not apply it in that setting. No longitudinal data structures other than the growth curve were considered.

Finally, Zeger and Liang [5] used generalized estimating equations (GEE) to model the correlation among repeated observations for a given subject and applied their methodology to dichotomous outcomes using a logistic regression model. The latter approach, however, depends on the repeated observation of the outcome variable, and, therefore, does not apply directly to the discrimination problem where the outcome occurs once and classification to a population is done after the last time of observation.

The most important limitation in the use of the linear discriminant model for longitudinal data is that the model is only applicable to fairly balanced data, an increasingly exceptional situation in longitudinal studies. Therefore, an approach is needed that does not rely on complete observations over time. The class of linear and non-linear mixed effects models, then, becomes an appropriate point of departure.

In Section 2, we summarize briefly the relevant developments in linear and non-linear mixed effects modelling methods. In Section 3, we extend the framework of classical discriminant analysis to the longitudinal setting, elaborate four models based on the variance structure of the longitudinal observations, and briefly discuss options for parameter estimation in these models, and approaches to evaluation of classification performance. We illustrate the proposed longitudinal discriminant method in Section 4 using data from Santiago, Chile on the subunit beta measured in women with normal and abnormal pregnancy outcomes.

2 Linear and Non-Linear Mixed Effects Models

The Linear Mixed Effects Model

Confronting similar limitations in multivariate linear models, Laird and Ware (1982) proposed a general linear mixed effects model to analyze longitudinal data, that recognizes the relationship between serial observations collected on the same subject. In their article, Laird and Ware proposed a combination of empirical Bayes and maximum likelihood methods via the EM algorithm to estimate the parameters of the mixed effects model. These models are based on earlier mixed effects models proposed by Harville (1977).

These models allow control in a situation with considerable variation among subjects both in the number and timing of observations and in the structure of the covariance matrix. Specifically, the Laird-Ware model makes it possible to estimate population curves, in addition to subject-specific curves, and, therefore, to discriminate with either single or multiple observations. This reduces significantly the total number of parameters to be estimated in the model and increases the reliability of the process of estimation. The model for the $n_i \times 1$ response vector $\boldsymbol{y}_i$ of the ith individual can be formulated as

$$\boldsymbol{y}_i = \boldsymbol{X}_i \boldsymbol{\alpha} + \boldsymbol{Z}_i \boldsymbol{b}_i + \boldsymbol{e}_i \tag{1}$$

where $\boldsymbol{\alpha}$ denotes a $p \times 1$ vector of unknown population parameters associated with the known $n_i \times p$ design matrix $\boldsymbol{X}_i$, and where $\boldsymbol{b}_i$ denotes a $k \times 1$ vector of unknown individual effects associated with the known design matrix $n_i \times k$. The error $\boldsymbol{e}_i$ is distributed as $N\left(0, \boldsymbol{W}_i\right)$, where $\boldsymbol{W}_i$ is an $n_i \times n_i$ positive-definite covariance matrix that only depends on i for its dimension. The vector $\boldsymbol{b}_i$ is distributed as $N\left(0, \boldsymbol{B}\right)$, independent of each other and of the $\boldsymbol{e}_i$, where $\boldsymbol{B}$ is a $k \times k$ positive definite covariance matrix. The marginal distribution of the response vector $\boldsymbol{y}_i$ is normal with mean $\boldsymbol{X}_i\boldsymbol{\alpha}$ and covariance matrix $\boldsymbol{V}_i = \boldsymbol{Z}_i \boldsymbol{B} \boldsymbol{Z}_i + \boldsymbol{W}_i$.

In a slightly different approach, Jennrich and Schluchter (1986) proposed a Newton-Raphson algorithm that allows more diversity in the covariance structure for the mixed effects model. Jones (1987, 1993) proposed the use of the Kalman Filter to estimate the parameters of these models in the presence of more complex serial correlation structures in the errors.

Non-linear Mixed Effects Models

Two very relevant extensions to non-linear mixed effects models have been proposed by Lindstrom and Bates(1990), Hirst, Zerbe et al (1991), and Vonesh and Carter (1992). In the first, the non-linear component of the model is reduced to the population curve, where the random effect is linear in the scale of the response. The model for the response vector can be represented as

$$\boldsymbol{y}_i = \boldsymbol{\mu}\left(\boldsymbol{X}_i, \boldsymbol{\alpha}\right) + \boldsymbol{Z}_i \boldsymbol{b}_i + \boldsymbol{e}_i, \tag{2}$$

with the usual structure in the distribution of $\boldsymbol{b}_i$ and $\boldsymbol{e}_i$. The marginal distribution of the response vector $\boldsymbol{y}_i$ is normal with mean $\boldsymbol{\mu}\left(\boldsymbol{X}_i, \boldsymbol{\alpha}\right)$ and covariance matrix $\boldsymbol{V}_i$ as in the linear model. A special case, that is relevant to our example, occurs when $\boldsymbol{Z}_i$ depends not only on the covariates $\boldsymbol{X}_i$, but also is a function of the parameters $\boldsymbol{\alpha}$.

The second extension is a more general model in which the random effect is also part of the non-linear component of the model. The model can be represented as

$$\boldsymbol{y}_i = \boldsymbol{\mu}\left(\boldsymbol{X}_i, \boldsymbol{\alpha}, \boldsymbol{Z}_i, \boldsymbol{b}_i\right) + \boldsymbol{e}_i \tag{3}$$

In this model, the marginal distribution of $\boldsymbol{y}_i$ can be difficult to find even in the case where the conditional distribution of $\boldsymbol{y}_i$ given $\boldsymbol{b}_i$ is normal and the marginal distribution of $\boldsymbol{b}_i$ is normal.

Another extension of these models is the generalized linear mixed effects model defined as

$$\boldsymbol{y}_i = \mu\left(\boldsymbol{X}_i \alpha + \boldsymbol{Z}_i \boldsymbol{b}_i\right) + \boldsymbol{e}_i$$

where y_{ij} given $\boldsymbol{b}_i$ has distribution in the exponential family (Zeger and Liang, 1986).

3 Longitudinal Discriminant Analysis

In this section we develop our longitudinal discriminant analysis approach combining the principles of the linear discriminant function and those of mixed effects models.

General Framework

Consider g populations or groups $G_1, G_2, \ldots, G_g$ $(g \geq 2)$, and suppose that associated with each one these populations there is a probability density function for the response vector $\boldsymbol{y}$, denoted as $f_1, f_2, \ldots, f_g$. If an individual comes from the population G_j, the vector of responses $\boldsymbol{y}$ taken at arbitrary times $t'=(t_1, t_2, \ldots, t_n)$ has probability density function $f\left(\boldsymbol{y}; \phi_j\left(t\right)\right)$, where the set of parameters associated with this distribution, $\phi_j\left(t\right)$, depend on t. Let $\boldsymbol{\mu}_j\left(t\right)$ and $\boldsymbol{V}_j\left(t\right)$ be the mean and the variance of $\boldsymbol{y}$ in population G_j. In this longitudinal discriminant analysis the goal is to allocate an individual into one of g groups on the basis of the observations $\boldsymbol{y}$, the time of these observations t, and the distribution of $\boldsymbol{y}$ in the g groups. In medical studies the times of observation are commonly arbitrary and unequally spaced, and the number of observations collected varies among subjects.

If we assume that $\pi_1, \pi_2, \ldots, \pi_g$ are the prior probabilities of group membership, the optimal rule for allocation classifies y to group G_k if

$$\log \pi_k + \log f\left(\boldsymbol{y}; \phi_k\left(t\right)\right) = \max_j \left\{\log \pi_j + \log f\left(\boldsymbol{y}\phi_j\left(t\right)\right)\right\}, j = 1, \ldots, g \tag{4}$$

where $f\left(\boldsymbol{y}; \phi_j(t)\right)\pi_j$ is the posterior distribution of membership in group j.

The densities $f\left(\boldsymbol{y}; \phi_j\left(t\right)\right)$ can be assumed to be the densities of a normal distribution with mean $\boldsymbol{\mu}_j\left(t\right)$ and variance $\boldsymbol{V}_j\left(t\right)$ if the observations $\boldsymbol{y}$ are generated by models (1) or (2) in which the subject specific effect $\boldsymbol{b}_i$ and the errors $\boldsymbol{e}_i$ are independent and distributed normal with mean zero and covariance matrix as specified before. The densities are not necessarily normal if the model from which the data were generated is nonlinear in the random effects $\boldsymbol{b}_i$, even in the case where these random parameters are normally distributed. The resulting marginal distribution of the response vector $\boldsymbol{y}$ is commonly unknown and difficult to find analytically. In such cases, the allocation rule (4) can not be used and an alternative procedure to the likelihood based method must be proposed.

In classical discriminant analysis the Mahalanobis distance plays a central role in both the conceptual framework, and the allocation rules. The Mahalanobis distance between the response vector $\boldsymbol{y}$ and the mean of the distribution of population G_j with respect to $\boldsymbol{V}_j(t)$ is

$$D_j^*\left(\boldsymbol{y}, t\right) = \left(\boldsymbol{y} - \boldsymbol{\mu}_j\left(t\right)\right)' \boldsymbol{V}_j(t)^{-1}\left(\boldsymbol{y} - \boldsymbol{\mu}_j\left(t\right)\right) + \log\left|\boldsymbol{V}_j\left(t\right)\right|. \tag{5}$$

Having defined this Mahalanobis distance, we can extend the allocation rule (4) to allocate $\boldsymbol{y}$ to group G_k if $\lambda_{jk}\left(\boldsymbol{y}; t\right) \leq 0$ for $j = 1, \ldots, g$ and $j \neq k$, where

$$\lambda_{jk}\left(\boldsymbol{y}; t\right) = D_k^*\left(\boldsymbol{y}, t\right) - D_j^*\left(\boldsymbol{y}, t\right) + 2\log\frac{\pi_j}{\pi_k} \tag{6}$$

The allocation rule (6) reduces to (4) when the density function $f\left(\boldsymbol{y}\phi_j(t)\right)$ belongs to the family of elliptic distribution with parameters $\boldsymbol{\mu}_j(t)$ and $\boldsymbol{V}_j(t)$, including the multivariate normal, multivariate Student's t, and Cauchy distribution.

There are many conceptual advantages to extending the Mahalanobis distance to expression (5). This allocation rule allows us to extend the optimal allocation rule to situations where the response vector is generated by models such as (3) where the distribution of $\boldsymbol{y}$ is generally unknown and difficult to compute, even in the case when the random component of the model $\boldsymbol{b}_i$ and $\boldsymbol{e}_i$ are normally distributed. The extended Mahalanobis distance is also consistent with the procedure of estimation since extended least squares is used as an estimation procedure when the method of maximum likelihood can not be applied such as in model (3).

If the distribution of the response vector $\boldsymbol{y}$ is $N\left(\boldsymbol{\mu}_j(t), \boldsymbol{V}_j(t)\right)$ in population G_j, which is the case when the data is generated by models (1) or (2) and the random component of the model $\boldsymbol{b}_i$ and $\boldsymbol{e}_i$ are normally distributed, then the maximum likelihood discriminant rule (4) allocates $\boldsymbol{y}$ to G_k if $\lambda_{jk}(\boldsymbol{y};t) \leq 0$ for $j = 1, 2, \ldots, g$ and $j \neq k$.

In summary, the principle of our approach is to take parameter estimates from subject-speciffic models and use them as estimators for population means and covariance matrices in classical linear and quadratic discriminant functions. For linear mixed effects models, these means and covariances derive from subject-specific observation vectors that are normally distributed. For non-linear mixed effects models, a linear approximation to the model residuals yields a marginal distribution of the subject-specific observation vector that is approximately normal [12], analogous to linear mixed effect models, with the marginal expectation evaluated at the random effect equal to zero. Some caution should be taken to note whether the linear approximation produces population mean esti- mates that are consistent with the subject-specific ones. In the following section, we describe the linear and quadratic discriminant functions that are possible under various parameter configurations for linear and non-linear mixed effects models, and an algorithm for parameter estimation.

Models

Four possible models are considered according to the form of the variance of $\boldsymbol{y}$ in the population G_j. The variance of $\boldsymbol{y}$ in the population G_j, $\boldsymbol{V}_j(t) = \boldsymbol{V}(t, \boldsymbol{\alpha}_j; \boldsymbol{\theta}_j)$, is a function of the mean population-specific parameters $\boldsymbol{\alpha}_j$ and the variance component $\boldsymbol{\theta}_j$. The general form of the variance can be written as

$$\boldsymbol{V}(t, \boldsymbol{\alpha}_j; \boldsymbol{\theta}_j) = \boldsymbol{Z}(t; \boldsymbol{\alpha}_j)\, \boldsymbol{B}(\boldsymbol{\theta}_j)\, \boldsymbol{Z}'(t; \boldsymbol{\alpha}_j) + \boldsymbol{W}(\boldsymbol{\theta}_j), \tag{7}$$

where $\boldsymbol{Z}(t; \boldsymbol{\alpha}_j)$ is a design matrix for the subject specific effect which is a function of the observation times of the response and the mean parameters $\boldsymbol{\alpha}_j$ for population G_j.

The Homoscedastic Model The homoscedastic model is obtained when $\boldsymbol{Z}(t; \boldsymbol{\alpha}_j) = \boldsymbol{Z}(t)$ does not depend on the mean parameters $\boldsymbol{\alpha}_j$ and the variance components are homogeneous, that is $\boldsymbol{\theta}_j = \boldsymbol{\theta}$ for $j = 1, 2, \ldots, g$. In this situation $\boldsymbol{V}_j(t) = \boldsymbol{V}(t)$ and the classification rule (6) with uniform priors allocates $\boldsymbol{y}$ to G_k if

$$D_k(\boldsymbol{y}, t) = (\boldsymbol{y} - \boldsymbol{\mu}_k(t))' \, \boldsymbol{V}(t)^{-1} (\boldsymbol{y} - \boldsymbol{\mu}_k(t))$$

is minimum among all groups. In particular, when $g = 2$ the rule allocates $\boldsymbol{y}$ to G_1 if

$$\boldsymbol{\beta}^{'} (\boldsymbol{y} - \overline{\boldsymbol{\mu}}(t)) \geq 0,$$

where $\boldsymbol{\beta} = \boldsymbol{V}(t)^{-1}(\boldsymbol{\mu}_1(t) - \boldsymbol{\mu}_2(t))$ and $\overline{\boldsymbol{\mu}}(t) = (\boldsymbol{\mu}_1(t) + \boldsymbol{\mu}_2(t))/2$, and to G_2 otherwise.

The *Mean*-Heteroscedastic Model The mean-heteroscedastic model consists of a model in which the design matrix $\boldsymbol{Z}(t;\boldsymbol{\alpha}_j)$ depends on the mean parameters $\boldsymbol{\alpha}_j$ but the variance component $\boldsymbol{\theta}_j = \boldsymbol{\theta}$ remain homogeneous among all population $j = 1, 2, \ldots, g$. In this case both the between and within subject variance $\boldsymbol{B}$ and $\boldsymbol{W}$ respectively are the same for all g populations. The dependence of $\boldsymbol{Z}\,(t;\boldsymbol{\alpha}_j)$ on $\boldsymbol{\alpha}_j$ may be due to the form of the underlying nonlinear model (2) or when the variance of $\boldsymbol{y}$ in model (**??**) can not be obtained analytically, the δ-method can be used to find an approximate variance of the form (7) where the design matrix $\boldsymbol{Z}$ is defined as the partial derivative of $\boldsymbol{\mu}\,(\boldsymbol{X}_i, \boldsymbol{\alpha_j}, \boldsymbol{Z}_i, \boldsymbol{b}_i)$ with respect to the random effects $\boldsymbol{b}_i$. In such cases the matrix $\boldsymbol{Z}_i$ differs across the g groups or populations.

The *Variance*-Heteroscedastic Model The component-heteroscedastic model consists of a model such as (1) or (2) in which the design matrix $\boldsymbol{Z}\,(t)$ does not depend on the population parameters $\boldsymbol{\alpha}_j$, but the variance component $\boldsymbol{\theta}_j$ are different across all populations $j = 1, 2, \ldots, g$. This is the case where the between subject variance $\boldsymbol{B}_j$ is different among the groups, or the within subject variance $\boldsymbol{W}_j$ varies among the g populations, or both covariance matrices are different.

The *Full*-Heteroscedastic Model The full-heteroscedastic model consists of a model such as (1) or (2) where the design matrix $\boldsymbol{Z}\,(t)$ depend on the population parameters $\boldsymbol{\alpha}_j$, and the variance component $\boldsymbol{\theta}_j$ are different across all populations $j = 1, 2, \ldots, g$.

Parameter Estimation

Population parameters and variance components in a linear mixed effects models under normal theory can be estimated using various procedures, including the EM algorithm proposed by Laird and Ware (**?**). In the context of a nonlinear mixed effects model, all homoscedastic models, all *mean*-heteroscedastic models, and some very special cases of the *variance*- and *fully*-heteroscedastic models can be fit using the algorithm of Lindstrom and Bates (**?**) and, therefore, using currently available computer software (**?**). Most *variance*- and *fully*-heteroscedastic models require some modification of the Lindstrom and Bates (**?**) algorithm to estimate the population parameters, $\boldsymbol{\alpha}_j$, and the variance components, $\boldsymbol{\theta}_j$ for each group or population.

We present here an algorithm for obtaining the parameter estimates for all models described above, with a slight modification of the Lindstrom and Bates (**?**) algorithm, and including the Laird and Ware (**?**) and the Jennrich and Schluchter (**?**) algorithms for linear mixed effects models. It is important to highlight that the vectors $\boldsymbol{\alpha}_j$ and $\boldsymbol{\theta}_j$ corresponding to the parameters of group j may contain elements in common with the vector of parameters of other groups. Therefore, for the purpose of describing parameter estimation, we define $\boldsymbol{\alpha}$ and $\boldsymbol{\theta}$ as $q \times 1$ and $s \times 1$ vectors containing the distinct parameters in $\boldsymbol{\alpha}_j$ and $\boldsymbol{\theta}_j$, respectively, across the groups. We introduce $\boldsymbol{A}_i$ and $\boldsymbol{C}_i$ as $p \times q$ and $t \times s$ design matrices such that $\boldsymbol{A}_i\boldsymbol{\alpha} = \boldsymbol{\alpha}_j$ and $\boldsymbol{C}_i\boldsymbol{\theta} = \boldsymbol{\theta}_j$ if subject i belongs to group j; where p and t are the dimensions of the vectors $\boldsymbol{\alpha}_j$ and $\boldsymbol{\theta}_j$, respectively.

The nonlinear mixed effects model for the ith subject is, then,

$$\boldsymbol{y}_i = \mu\,(\boldsymbol{X}_i, \boldsymbol{A}_i\boldsymbol{\alpha}, \boldsymbol{b}_i) + \boldsymbol{e}_i$$

and the marginal variance of $\boldsymbol{y}_i$ in the most general case can be approximated by (7). The linearized version of this model is

$$\tilde{\boldsymbol{y}}_i = \tilde{\boldsymbol{X}}_i \boldsymbol{A}_i \boldsymbol{\alpha} + \tilde{\boldsymbol{Z}}_i \boldsymbol{b}_i + \boldsymbol{e}_i \tag{8}$$

where $\tilde{\boldsymbol{y}}_i = \boldsymbol{y}_i - \mu\left(\boldsymbol{X}_i, \boldsymbol{A}_i\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{b}}_i\right) + \tilde{\boldsymbol{X}}_i \boldsymbol{A}_i \widehat{\boldsymbol{\alpha}} + \tilde{\boldsymbol{Z}}_i \widehat{\boldsymbol{b}}_i$,

$$\tilde{\boldsymbol{X}}_i = \left(\left. \frac{\partial \mu\left(\boldsymbol{X}_i, \boldsymbol{A}_i\boldsymbol{\alpha}, \boldsymbol{b}_i\right)}{\partial \boldsymbol{\alpha}} \right|_{\boldsymbol{\alpha}=\widehat{\boldsymbol{\alpha}}, \boldsymbol{b}_i=\widehat{\boldsymbol{b}}_i} \right) \boldsymbol{A}_i^{'},$$

and $\tilde{\boldsymbol{Z}}_i$ is, similarly, the partial derivative of $\mu\left(\boldsymbol{X}_i, \boldsymbol{A}_i\boldsymbol{\alpha}, \boldsymbol{b}_i\right)$ with respect to $\boldsymbol{b}_i$ evaluated at the current values of the estimates $\widehat{\boldsymbol{\alpha}}$ and $\widehat{\boldsymbol{b}}_i$.

Given initial values for $\widehat{\boldsymbol{\alpha}} = \boldsymbol{\alpha}^0$, $\widehat{\boldsymbol{b}}_i = 0$ and $\widehat{\boldsymbol{\theta}} = \boldsymbol{\theta}^0$, the estimation algorithm takes the following steps:

1. Compute $\tilde{\boldsymbol{y}}_i$, $\tilde{\boldsymbol{X}}_i$, and $\tilde{\boldsymbol{Z}}_i$ as described above.
2. Update $\widehat{\boldsymbol{\alpha}}$ as
$$\widehat{\boldsymbol{\alpha}} = \left(\sum_{i=1}^{m} \boldsymbol{A}_i^{'} \tilde{\boldsymbol{X}}_i^{'} \widehat{\boldsymbol{V}}_i^{-1} \tilde{\boldsymbol{X}}_i \boldsymbol{A}_i \right)^{-1} \sum_{i=1}^{m} \boldsymbol{A}_i^{'} \tilde{\boldsymbol{X}}_i^{'} \widehat{\boldsymbol{V}}_i^{-1} \tilde{\boldsymbol{y}}_i$$
where $\widehat{\boldsymbol{V}}_i = \boldsymbol{V}_i(\boldsymbol{C}_i\widehat{\boldsymbol{\theta}})$. Following Lindstrom and Bates (**?**), $\tilde{\boldsymbol{y}}_i \overset{\cdot}{\sim} N\left(\tilde{\boldsymbol{X}}_i \boldsymbol{A}_i \boldsymbol{\alpha}, \boldsymbol{V}_i\right)$, and $\widehat{\boldsymbol{\alpha}}$ is the MLE based on the pseudo-likelihood of $\tilde{\boldsymbol{y}}_i$.
3. Update $\widehat{\boldsymbol{b}}_i$ as
$$\widehat{\boldsymbol{b}}_i = \widehat{\boldsymbol{B}}_i \tilde{\boldsymbol{Z}}_i \widehat{\boldsymbol{V}}_i^{-1} \tilde{\boldsymbol{r}}_i$$
where $\widehat{\boldsymbol{B}}_i = \boldsymbol{B}(\boldsymbol{C}_i\widehat{\boldsymbol{\theta}})$ and $\tilde{\boldsymbol{r}}_i = \tilde{\boldsymbol{y}}_i - \tilde{\boldsymbol{X}}_i \boldsymbol{A}_i \widehat{\boldsymbol{\alpha}}$. Again, following Lindstrom and Bates (**?**), $\widehat{\boldsymbol{b}}_i$ is the conditional expectation, $E\left\{\boldsymbol{b}_i \mid \tilde{\boldsymbol{y}}_i, \widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\theta}}\right\}$.
4. Update the r-th element of $\widehat{\boldsymbol{\theta}}$, $\widehat{\theta}_r$, by solving the estimating equation (**?**)
$$\sum_{i=1}^{m} \operatorname{tr}\left\{ \boldsymbol{V}_i^{-1} \left(\tilde{\boldsymbol{r}}_i \tilde{\boldsymbol{r}}_i' - \boldsymbol{V}_i\right) \boldsymbol{V}_i^{-1} \dot{\boldsymbol{V}}_{ir} \right\} = 0$$
where $\dot{\boldsymbol{V}}_{ir} = \partial \boldsymbol{V}_i / \partial \theta_r$. Alternatively, using a REML-type estimating equation we replace $\tilde{\boldsymbol{r}}_i \tilde{\boldsymbol{r}}_i' - \boldsymbol{V}_i$ in the previous equation by
$$\tilde{\boldsymbol{r}}_i \tilde{\boldsymbol{r}}_i' - \boldsymbol{V}_i + \tilde{\boldsymbol{X}}_i \boldsymbol{A}_i \left(\sum_{l=1}^{m} \boldsymbol{A}_l^{'} \tilde{\boldsymbol{X}}_l^{'} \boldsymbol{V}_l^{-1} \tilde{\boldsymbol{X}}_l \boldsymbol{A}_l \right)^{-1} \boldsymbol{A}_i^{'} \tilde{\boldsymbol{X}}_i^{'}.$$
5. Return to Step 1. until convergence is obtained.

Evaluating the Longitudinal Discriminant Method

For the classical linear discriminant function the percent of observations classified correctly into their parent populations has been the most common means by which to evaluate how well the function performs. Unbiased estimates of these classification probabilities can be obtained using jackknife estimates when the estimated function is applied to the training data (Lachenbruch, 1965) or using cross-validation when a hold-out sample is available. Such estimates can also be applied in the longitudinal data setting, but additional features of classification arise in this context and these will be the focus of this section.

Cumulative Information over Time The cumulative information provided by observations taken over time has the potential for increasing the sensitivity and specificity of the longitudinal discriminant method. Receiver operating characteristic (ROC) curves (Hanley and McNeil, 1982) can be produced which show the classification results as a function of the time since beginning of observation. The different curves produced based on cumulative data can be compared based on the areas under each ROC curve (Hanley and McNeil, 1983).

Timing of observations In some cases, particular time intervals are more informative of the longitudinal process than others. In this case, the evaluation of the discriminant function may take an interval-specific form. The ROC curves would then show the tradeoff between sensitivity and specificity for different intervals of time in which data are collected. The approach of Zeger and Liang (1986) could additionally be used to model correct classification as a binary outcome, longitudinally, as a function of the time of observation.

Confounding of Number of Observations with Time ROC curves showing the cumulative effects of the number of observations on classification can also be obtained. However, the number of observations can be confounded by the time or timing of observation, e.g. some individuals may have many early observations and no late ones or vice versa, or some may have more observations during an informative period of the longitudinal process. This evaluation, then, needs to be adjusted for the time of observation. As suggested above, correct classification could be modeled as a function of both the number and time of observation.

Marginal Cost of Additional Observation vs. Gain in Classification As indicated above, additional observations are expected to lead to gains in correct classification. These gains can often be accompanied by substantial economic costs, however. In that case, the cost of taking an additional longitudinal measurement needs to be weighed against the expected cost of reduced classification when no additional measurement is taken. Information would continue to be collected as long as its cost did not exceed the expected cost of not obtaining the information. Using this approach, the number and timing of measurements could be optimized in future applications of the discriminant function.

In summary, the longitudinal setting leads to additional possibilities for and difficulties with evaluation of the discriminant function and the measurement process itself. The first of these steps for evaluation, looking at the effect of cumulative information over time, will be illustrated with the example below.

4 An Example

It is well known in obstetrics that, among other clinical variables, the sub-unit beta shows dramatic changes in women during pregnancy. It has been established also that values of the sub-unit beta are different in women who have normal pregnancies with terminal deliveries than in women who have spontaneous abortions or other types of adverse pregnancy outcomes. This association has made it possible to classify, with some uncertainty, the outcome of pregnancy.

In a follow-up study of 150 young women, representing 150 different pregnancies, over a period of 2 years in a private obstetrics clinic in Santiago, Chile, the values of the sub-unit beta were measured during the first 80 days of gestational age. The women were classified as normal, if they had a normal terminal delivery, or as abnormal if they had any complication resulting in a nonterminal delivery and loss of the fetus.

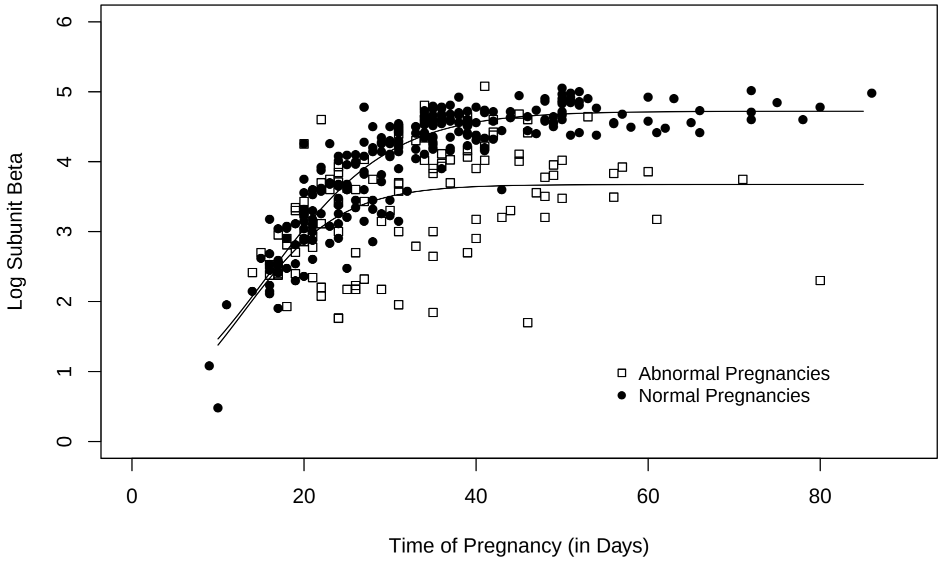


Fig. 1. Logistic curve for the subunit beta

Mean values of the log sub-unit beta by day of gestational age show a nonlinear relationship, with a threshold after 50 days of pregnancy. The logistic response curve model shown in Figure **??** is a reasonable alternative to describe the changes in the sub-unit beta in the log scale across the days of pregnancy. The proposed model is

$$y_{ijk} = \frac{\alpha_{k1} + b_{ik}}{1 + \alpha_{k2} \exp\{-\alpha_{k3} t_{ij}\}} + e_{ijk}$$

where $k = 1$ for the normal pregnancy group and $k = 2$ for the abnormal pregnancy group. The random components of this model are assume to have normal distribution , that is $b_{ik} \sim N(0, B_k)$ and $e_{ijk} \sim N\left(0, \sigma_k^2\right)$, where the e_{ijk}'s are assumed to be mutually independent within and between subjects and independent of the b_{ik}'s.

The set of parameters $\boldsymbol{\alpha}_1' = (\alpha_{11}, \alpha_{12}, \alpha_{13})$ and $\boldsymbol{\alpha}_2' = (\alpha_{21}, \alpha_{22}, \alpha_{23})$ represent the population parameters of the logistic curve for the normal pregnancy group and for the abnormal pregnancy group respectively. An overall test for the hypothesis $H_0 : \boldsymbol{\alpha}_1 = \boldsymbol{\alpha}_2$ will provide an idea of the total discriminatory power of this model for the data collected. A test of hypothesis for

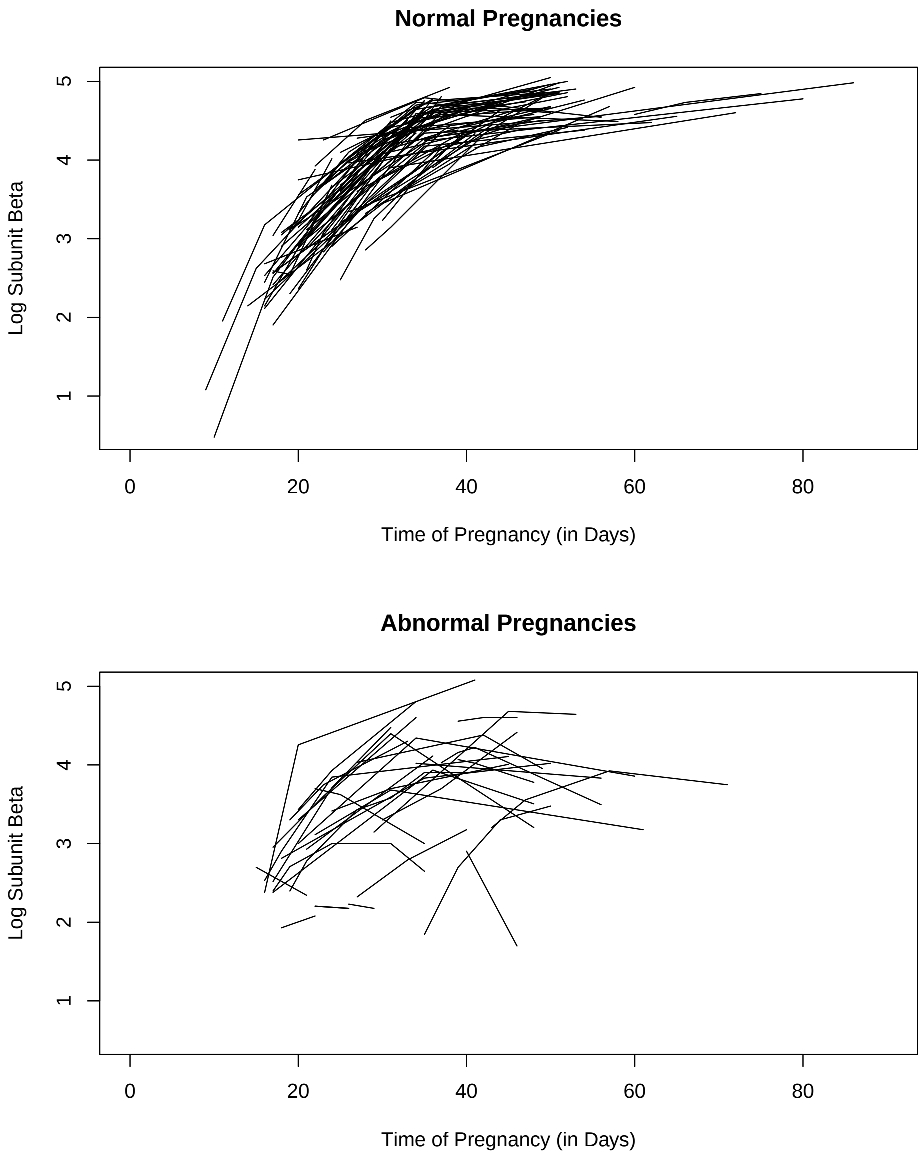


Fig. 2. Time profiles for randomily selected normal and Abnormal subjects

individual parameters will provide information about the contribution of each parameter to the discrimination process.

The marginal distribution of the response vector $\boldsymbol{y}_i$ in the k th group is

$$\boldsymbol{y}_i \sim N\left(\boldsymbol{\mu}_k(t_i), B_k \boldsymbol{z}_{ik}\boldsymbol{z}'_{ik} + \sigma_k^2 \boldsymbol{I}_i\right), \tag{9}$$

where the mean vector $\boldsymbol{\mu}_k(t_i)$ has elemets

$$\mu_k(t_{ij}) = \frac{\alpha_{k1}}{1+\alpha_{k2}\exp\{-\alpha_{k3}t_{ij}\}}, \tag{10}$$

and represent the population curve at times t_{ij}, $\boldsymbol{I}_i$ is a $n_i \times n_i$ identity matrix, and the vector $\boldsymbol{z}_{ik}$ has elements

$$z_{ijk} = \frac{1}{1+\alpha_{k2}\exp\{-\alpha_{k3}t_{ij}\}}$$

which depend on the values of the unknown population parameters α_{k2} and α_{k3}. The linearized version of model (9) and (10) is

$$\boldsymbol{y}_{ik} = \boldsymbol{X}_{ik}\boldsymbol{\alpha}_k + \boldsymbol{z}_{ik}b_{ik} + \boldsymbol{e}_{ik},$$

where

$$\boldsymbol{X}_{ik} = \frac{\partial \mu_k(t_{ij})}{\partial \boldsymbol{\alpha}_k} = \begin{pmatrix} z_{ijk} \\ -\exp\{-\alpha_{k3}t_{ij}\}\, z_{ijk}\mu_k(t_{ij}) \\ \alpha_{k2}t_{ij}\exp\{-\alpha_{k3}t_{ij}\}\, z_{ijk}\mu_k(t_{ij}) \end{pmatrix},$$

and $\boldsymbol{z}_{ik}$ as before.

Model (9) and (10) correspond to a *full* -heteroscedastic model. A *mean*-Heteroscedastic model can be obtained from expression (**??**.1.Model-1) by introducing the restrictions $B_1 = B_2$ and $\sigma_1^2 = \sigma_2^2$, and the *Variance*-Heteroscedastic model from expression (10) by introducing the restrictions $\alpha_{12} = \alpha_{22}$ and $\alpha_{13} = \alpha_{23}$. The homoscedastic model is obtained by introducing the two set of restrictions showed above.

The results of fitting these four models are shown in Table 1. From this Table we can concluded that the full-heteroscedastic model is the best to discriminate among normal and non-terminal deliveries.

Table 1. Summary of Model Fitting

Model	df	$-2\log L$	AIC	χ^2	$p-$value
Homoscedastic	6	-203.8	-191.8	108.8	< 0.01
Mean-Heteroscedastic	8	-231.7	-215.7	80.9	< 0.01
Variance-Heteroscedastic	8	-299.3	-283.3	13.3	< 0.01
Full-Heteroscedastic	10	-312.6	-292.6	-	-

The estimates of the four models are shown in Table 2. We can conclude by examining the estimates of the variance component of the *full*-heteroscedastic that the abnormal group presents

Table 2. Estimates of the Discriminant Models

	Estimates of Model			
	Homoscedastic	Heteroscedastic		
Parameter		*Mean*	*Variance*	*Full*
α_{11}	4.656	4.740	4.689	4.722
α_{12}	8.380	9.017	8.582	8.914
α_{13}	0.144	0.140	0.141	0.139
α_{21}	3.916	3.680	3.956	3.674
α_{22}	-	8.103	-	9.776
α_{23}	-	1.089	-	0.177
B_1	0.158	0.167	0.031	0.033
B_2	-	-	0.609	0.540
σ_1^2	0.132	0.115	0.092	0.090
σ_2^2	-	-	0.192	0.162

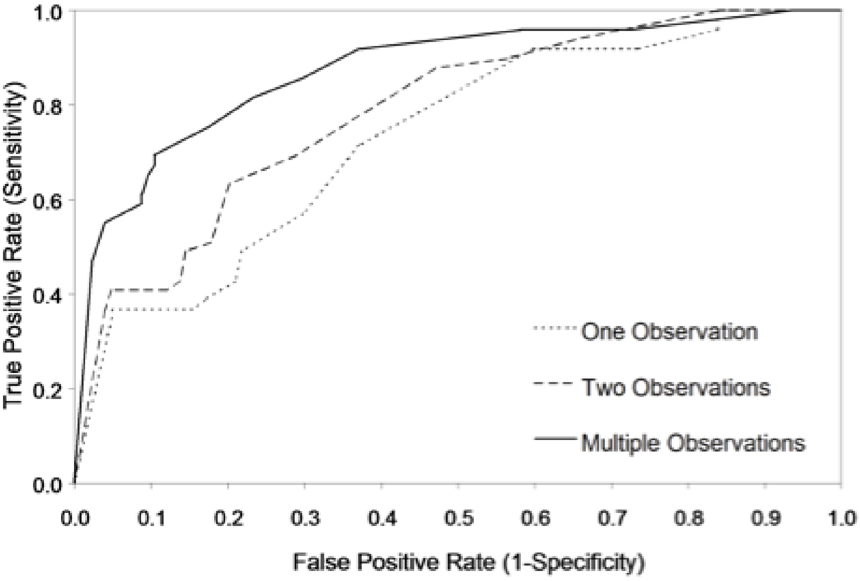


Fig. 3. ROC curve for discriminant model

a significantly more variability among the subject specific curves than the normal group and the model is less precise.

The ROC curves for the full-heteroscedastic model are presented in Figure 3. Three curves are presented which show the changes in sensitivity and specificity using only the first, only the first and second, and using all of the available beta-subunit observations on each woman. While only small gains in sensitivity and specificity are seen with one vs. two subunit beta measures on a woman, there is a greater effect of using multiple observations of the subunit beta to improve the sensitivity and specificity for predicting an abnormal pregnancy outcome in this population of women.

5 Discussion

The principal advantage of the discriminant model for unbalanced repeated measures proposed is the ability to use all of the information for classifying subjects over time, regardless of the number or the time of the observations. As we have shown in our example the predictive capability of a model using all the information can be increased significantly. This methodology is more appropriate for clinical practice where the number and times of observation are often arbitrary and depend on the progression of the patient.

The use of this method in clinical practice solves an important problem but opens new ones. For example, if observations have an associated economic cost or a medical cost for the patient, such as with mammography, the optimal time and number of observation should also be investigated.

Bibliography

[1] Albert, A. (1983). Discriminant Analysis Based on Multivariate Response Curves: A Descriptive Approach to Dynamic Allocation. *Statistics in Medicine* 2, 95-106.

[2] Albert, A., Chapelle J. P., Heusghem, C., Kulbertus, H. E., and Harris, E. K. (1982). Evaluation of Risk Using Serial Laboratory Data in Acute Myocardial Infarction, in C Heisghem, A Albert and ES Benson (Eds.) *Advanced Interpretation of Clinical Laboratory Data*, New York: Marcel Dekker.

[3] Albert, J. M. and Kshirsagar, A. M. (1993). The Reduced-Rank Growth Curve Model for Discriminant Analysis of Longitudinal Data. *Australian Journal of Statistics*, 35:345-357.

[4] Azen, S. P. and Afifi, A. A. (1972). Two Models for Assessing Prognosis on the Basis of Successive Observations. *Mathematical Biosciences*, 14,169-176.

[5] Hanley, J. A. and McNeil, B. J. (1982). The Meaning and Use of the Area Under a Receiver Operating Characteristic (ROC) Curve. *Radiology*, 143:29-36.

[6] Hanley, J. A. and McNeil, B. J. (1983). A Method of Comparing the Areas Under Receiver Operating Characteristics Curves Derived from the Same Cases. *Radiology*, 148:839-843.

[7] Harville, D. A. (1977). Maximum Likelihood Approaches to Variance Component Estimation and to Related Problems. *Journal of the American Statistical Association* 72,320-340.

[8] Hirst, K., Zerbe, G. O., Boyle, D. W. and Wilkening, R. B. (1991). On nonlinear random effects models for repeated measurements. *Communications in Statistics B, Simulation and Computation* 20: 463-478.

[9] Jennrich, R. I. and Schluchter, M. D. (1986). Unbalanced Repeated Measures Models with Structural Covariance Matrices. *Biometrics* 42, 805-820.

[10] Jones, R. H. (1987). Serial correlation in unbalanced mixed models. Invited paper 23.2, 46th session of the ISI.

[11] Jones, R. H. (1993) Longitudinal Data with Serial Correlation: A State-space Approach. New York:Chapman and Hall.

[12] Laird, N. M. and Ware, JH (1982). Random-Effects Models for Longitudinal Data. *Biometrics* 38, 963-974.

[13] Lindstrom, M. J. and Bates, D. M. (1990). Nonlinear Random Effects Models for Repeated Measures Data. *Biometrics* 46, 673-687.

[14] Lachenbruch, P. A. (1965). Estimation of Error Rates in Discriminant Analysis. Ph.D. dissertation. University of California at Los Angeles.

[15] Lachenbruch, P. A. (1975). *Discriminant Analysis* . New York: Hafner Press.

[16] Vonesh, E. F. and Carter, R. L. (1992). Mixed-effects nonlinear regression for unbalanced repeated measures. *Biometrics* 48, 1-18.

[17] Zeger, S. L. and Liang, K. (1986).Longitudinal Data Analysis for Discrete and Continuous Outcomes. *Biometrics*, 42:121-130.

Article 2.2

Nonlinear random effects model for multivariate responses with missing data

Guillermo Marshall, Rolando de la Cruz-Mesía, Anna E. Barón, James H. Rutledge, and Gary O. Zerbe

Pontificia Universidad Católica de Chile and
University of Colorado HSC, Denver, U.S.A.

Abstract. The use of random-effects models for the analysis of longitudinal data with missing responses has been discussed by several authors. In this paper, we extend the nonlinear random-effects model for a single response to the case of multiple responses, allowing for arbitrary patterns of observed and missing data. Parameters for this model are estimated via the EM algorithm. The set of equations for this estimation procedure is derived and these are appropriately modified to deal with missing data. The methodology is illustrated with an example using data coming from a study involving 161 pregnant women presenting to a private obstetrics clinic in Santiago, Chile. [1]

1 Introduction

Analysis of longitudinal studies in which a single characteristics is measured at n different occasions has been considered by many authors, including Grizzle and Allen (1), Khatri (2), Potthoff and Roy (3), and Rao (4). They have all considered linear models where the timing of observations to be the same for all units and made no allowance for missing data. The case of arbitrary measurement times and variation in the number of observations for different individuals has been considered by Laird and Ware (6), who used a random effects approach. An extension to generalized linear models with multiple responses in which a random effects model for the covariance structure is assumed has been considered by Reinsel (7). He dealt with a complete and balanced design in which all individuals are measured at the same time points and they all have an equal number of observations (i.e., there is no allowance for missing data). In clinical trials dropouts are common, and it is also possible that units are measured at arbitrary times. In this case iterative techniques such as the EM or Newton-Raphson algorithms must be used to obtain maximum likelihood (ML) estimates.

Despite the popularity of single response models, multivariate versions of the linear random effects models have received scant treatment in the literature. Mickey *et al.* (8) presented methods for the analysis of multiple outcome variables collected longitudinally. Further summarization of the outcome variables in terms of their linear combination is investigated. Zucker *et al.* (9) made substantial improvements to Mickey's model. Shah *et al.* (10) extended the EM-type algorithm of Laird and Ware (6) to a bivariate setting for complete and incomplete data. They considered

[1] Marshall G, de la Cruz-Mesía R, Barón AE, Rutledge JH, Zerbe GO, Nonlinear random effects model for multivariate responses with missing data. *Statistics in Medicine* , 25 (16). 2817-2830.

two different structures for the covariance matrix of measurement error: uncorrelated error between responses and correlation of error terms at the same measurement time. More recently, Schafer and Yucel (5) developed new computational techniques for multivariate longitudinal data with missing data. Using a multivariate extension of a popular linear random effects model, they create multiple imputations of missing values for subsequent analyses by a straightforward and effective Markov chain Monte Carlo procedure. A new EM algorithm is used to estimate the parameters and converges more rapidly than traditional EM algorithms because it does not treat the random effects as missing data, but integrates them out of the likelihood function analytically. An implicit assumption of MAR is made however.

Approximate MLE of the population parameters for nonlinear random effects models was pioneered by Beal and Sheiner (11), and since then a number of algorithms have appeared for approximate ML, including Steimer *et al.* (12), Lindstrom and Bates (13), Hirst *et al.* (14), Beal and Sheiner (15), Vonesh and Carter (16), and Mentre and Gomeni (17). All of these algorithms are approximate in some way. For a summary see Beal and Sheiner (15), Wolfinger (18), Pinheiro and Bates (19), and Davidian and Giltinan (20). An EM algorithm for exact maximum likelihood estimation of a class of nonlinear random effects models is given by Walker (21).

In this article we consider the situation in which a given individual has p different characteristics measured at each of n_i different occasions under known, possibly different experimental conditions. We begin by assuming that all p characteristics are observed at each occasion, although the number and timing of observations may differ from individual to individual; that is, the data are complete in characteristics but unbalanced in number of observations per experimental unit. We show that for this case, the EM algorithm presented by Lindstrom and Bates (13) can be easily extended by treating the random error terms as missing data. We then consider the more difficult case where only a subset of the p characteristics may be observed at any occasion; in this setting both the random effects and any unobserved characteristics are considered to be "missing data" in our implementation of the EM algorithm.

Multivariate longitudinal data are characterized by multiple responses measured at multiple occasions for each subject. One such example concerns pregnant women. During pregnancy a variety of quantities or characteristics are measured at the prenatal examinations, in order to detect complications. In 161 pregnant women presenting to a private obstetrics clinic in Santiago, Chile, beta-subunit beta and estradiol were measured repeatedly over time for each woman. We use these values to predict normal versus abnormal pregnancy outcomes. As we will illustrate, these bivariate repeated measures can be easily modelled with the approach we propose.

2 Nonlinear Mixed Effects Model for Multivariate Longitudinal Data

Let $\boldsymbol{Y}_i$ be an $n_i \times p$ response matrix for the ith individual collected longitudinally, in which the rows represent n_i different times of observation and the columns the p different response measurements. This is,

$$\boldsymbol{Y}_i = \begin{pmatrix} y_{i11} & y_{i12} & \cdots & y_{i1p} \\ y_{i21} & y_{i22} & \cdots & y_{i2p} \\ \vdots & \vdots & \ddots & \vdots \\ y_{in_i1} & y_{in_i2} & \cdots & y_{in_ip} \end{pmatrix}$$

and let $\boldsymbol{E}_i$ be the $n_i \times p$ matrix of error terms associated with $\boldsymbol{Y}_i$. Introducing the vec operator, which strings out the columns of a matrix vertically, we can obtain $\boldsymbol{y}_i = \mathrm{vec}(\boldsymbol{Y}_i) = \left(\boldsymbol{y}'_{i1}, \boldsymbol{y}'_{i2}, \ldots, \boldsymbol{y}'_{ip}\right)'$ the $pn_i \times 1$ vector of outcomes and $\boldsymbol{\epsilon}_i = \mathrm{vec}(\boldsymbol{E}_i)$ is similar for the error term.

We assume a model for individual i at time j for the kth response to be of the form

$$y_{ijk} = f_k(\boldsymbol{\eta}_i, \boldsymbol{x}_{ij}) + \epsilon_{ijk}$$

where f_k is a nonlinear function of the parameter vector $\boldsymbol{\eta}_i$ and the covariate vector $\boldsymbol{x}_{ij}$, and ϵ_{ijk} is the error term in the model. The parameter vector $\boldsymbol{\eta}_i$ can be incorporated into the model as $\boldsymbol{\eta}_i = \mathbf{A}_i\boldsymbol{\beta} + \boldsymbol{B}_i\boldsymbol{\beta}_i$, where $\boldsymbol{\beta}$ is a $q \times 1$ vector of fixed population parameters, $\boldsymbol{\beta}_i$ is an $r \times 1$ vector of individual random effects, and the matrices $\mathbf{A}_i$ and $\boldsymbol{B}_i$ are design matrices of size $s \times q$ and $s \times r$ for the fixed and random effects, respectively. We assume that $\boldsymbol{\beta}_i \sim \mathrm{MVN}(0, \boldsymbol{D}_{r\times r})$ and $\boldsymbol{\epsilon}_i \sim \mathrm{MVN}(0, \boldsymbol{R}_i)$ where $\boldsymbol{R}_i$ has dimensions $pn_i \times pn_i$. We will also assume that the error terms of different subjects are not correlated, that is $\mathrm{cov}(\boldsymbol{\epsilon}_i, \boldsymbol{\epsilon}_{i'}) = 0$, and the random effects and the error terms for the same subject are not correlated, that is, $\mathrm{cov}(\boldsymbol{\epsilon}_i, \boldsymbol{\beta}_i) = 0$.

We can write the model for the kth column of $\boldsymbol{Y}_i$ as

$$\boldsymbol{y}_i^k = \boldsymbol{\mu}_{ik}(\boldsymbol{\eta}_i, \boldsymbol{X}_i) + \boldsymbol{\epsilon}_i^k$$

where $\boldsymbol{y}_i^k = (y_{i1k}, \ldots, y_{in_ik})'$, $\boldsymbol{\mu}_{ik}(\boldsymbol{\eta}_i, \boldsymbol{X}_i) = (f_k(\boldsymbol{\eta}_i, \boldsymbol{x}_{i1}), \ldots, f_k(\boldsymbol{\eta}_i, \boldsymbol{x}_{in_i}))'$ and $\boldsymbol{\epsilon}_i^k = (\epsilon_{i1k}, \ldots, \epsilon_{in_ik})'$, the model for the jth row of $\boldsymbol{Y}_i$ can be written as

$$\boldsymbol{y}_{ij} = \boldsymbol{\mu}_i(\boldsymbol{\eta}_i, \boldsymbol{x}_{ij}) + \boldsymbol{\epsilon}_{ij}$$

where $\boldsymbol{y}_{ij} = (y_{ij1}, \ldots, y_{ijp})'$, $\boldsymbol{\mu}_i(\boldsymbol{\eta}_i, \boldsymbol{x}_{ij}) = (f_1(\boldsymbol{\eta}_i, \boldsymbol{x}_{ij}), \ldots, f_p(\boldsymbol{\eta}_i, \boldsymbol{x}_{ij}))'$ and $\boldsymbol{\epsilon}_{ij} = (\epsilon_{ij1}, \ldots, \epsilon_{ijp})'$ or the p responses for the ith individual as

$$\boldsymbol{y}_i = \boldsymbol{\mu}_i(\boldsymbol{\eta}_i, \boldsymbol{X}_i) + \boldsymbol{\epsilon}_i.$$

There are different forms in which to represent $\boldsymbol{R}_i$, the covariance matrix of $\boldsymbol{\epsilon}_i$, in order to reduce the number of parameters to be estimated. One of the natural representations of $\boldsymbol{R}_i$ is to apply to the multivariate responses the covariance matrix of the error term used in univariate responses, that is, $\boldsymbol{R}_i = \sigma^2\boldsymbol{I}_i$. If we represent $\boldsymbol{E}_{i[j]}$ as the jth row of the error term $\boldsymbol{E}_i$, and assume that $\boldsymbol{E}_{i[j]} \sim MVN(0, \boldsymbol{\Sigma}_{p\times p})$ and $\mathrm{cov}(\boldsymbol{E}_{i[j]}, \boldsymbol{E}_{i'[j']}) = 0$ for all $i = 1, \ldots, N$ and $j = 1, \ldots, n_i$ except when both $i = i'$ and $j = j'$, then $\boldsymbol{R}_i$ can be written as

$$\boldsymbol{R}_i = \boldsymbol{\Sigma} \otimes \boldsymbol{I}_i.$$

This assumption implies that error terms for different responses from the same subject and for observations measured at the same time have covariance structure $\boldsymbol{\Sigma}$. All other error terms are considered uncorrelated.

Using a Taylor series expansion about an initial guess $\boldsymbol{\eta}_i^0 = \mathbf{A}_i\boldsymbol{\beta}^0 + \boldsymbol{B}_i\boldsymbol{\beta}_i^0$, the model becomes

$$y_{ijk} - f_k(\boldsymbol{\eta}_i^0, \boldsymbol{x}_{ij}) + \dot{f}_k(\boldsymbol{\eta}_i^0, \boldsymbol{x}_{ij})'\boldsymbol{\eta}_i^0 = \dot{f}_k(\boldsymbol{\eta}_i^0, \boldsymbol{x}_{ij})'\boldsymbol{\eta}_i + \epsilon_{ijk}$$

where $\dot{f}_k$ are the partial derivatives of f_k with respect to $\boldsymbol{\eta}$. For convenience we can rewrite this model as

$$\tilde{y}_{ijk} = \tilde{\boldsymbol{x}}_{ijk}\boldsymbol{\beta} + \tilde{\boldsymbol{z}}_{ijk}\boldsymbol{\beta}_i + \epsilon_{ijk}$$

where $\tilde{y}_{ijk} = y_{ijk} - f_k(\boldsymbol{\eta}_i^0, \boldsymbol{x}_{ij}) + \dot{f}_k(\boldsymbol{\eta}_i^0, \boldsymbol{x}_{ij})'\boldsymbol{\eta}_i^0$, $\tilde{\boldsymbol{x}}_{ijk} = \dot{f}_k(\boldsymbol{\eta}_i^0, \boldsymbol{x}_{ij})'\mathbf{A}_i$ is a $1 \times q$ vector , and $\tilde{\boldsymbol{z}}_{ijk} = \dot{f}_k(\boldsymbol{\eta}_i^0, \boldsymbol{x}_{ij})'\boldsymbol{B}_i$ is a $1 \times r$ vector.

The model for the super vector of all *pseudo*-responses for the ith subject is

$$\tilde{\boldsymbol{y}}_i = \tilde{\boldsymbol{X}}_i\boldsymbol{\beta} + \tilde{\boldsymbol{Z}}_i\boldsymbol{\beta}_i + \boldsymbol{\epsilon}_i$$

where $\tilde{\boldsymbol{X}}_i$ is a $pn_i \times q$ matrix with rows made up of $\tilde{\boldsymbol{x}}_{ijk}$ and $\tilde{\boldsymbol{Z}}_i$ is a $pn_i \times r$ matrix with rows made up of $\tilde{\boldsymbol{z}}_{ijk}$.

Computation via the EM Algorithm

We compute the parameter estimates for the model using the EM algorithm. The complete data for estimating $\boldsymbol{\Sigma}$, $\boldsymbol{D}$, and $\boldsymbol{\beta}$ are $\tilde{\boldsymbol{y}}_i$, $\boldsymbol{\beta}_i$ and $\boldsymbol{\epsilon}_i$. The sufficient statistic for $\boldsymbol{\Sigma}$ is $\sum_{i=1}^N \boldsymbol{E}_i'\boldsymbol{E}_i$, where $\boldsymbol{E}_i$ is the corresponding error term of $\boldsymbol{Y}_i$ in matrix form, that is $\boldsymbol{\epsilon}_i = \text{vec}(\boldsymbol{E}_i)$. The sufficient statistic for $\boldsymbol{D}$ is $\sum_{i=1}^N \boldsymbol{\beta}_i\boldsymbol{\beta}_i'$.

It is because $\boldsymbol{\beta}_i$ and $\boldsymbol{\epsilon}_i$ are not observed that we use the EM algorithm. The E-step starts with

$$\boldsymbol{\beta}^{(\nu)} = \left(\sum_{i=1}^N \tilde{\boldsymbol{X}}_i'\boldsymbol{W}_i^{(\nu)}\tilde{\boldsymbol{X}}_i\right)^{-1} \sum_{i=1}^N \tilde{\boldsymbol{X}}_i'\boldsymbol{W}_i^{(\nu)}\tilde{\boldsymbol{y}}_i$$

where $\boldsymbol{W}_i^{(\nu)} = \left(\tilde{\boldsymbol{Z}}_i\boldsymbol{D}^{(\nu)}\tilde{\boldsymbol{Z}}_i' + \boldsymbol{\Sigma}^{(\nu)} \otimes \boldsymbol{I}_i\right)^{-1}$.

The conditional expectation of $\boldsymbol{\beta}_i\boldsymbol{\beta}_i'$ given $\boldsymbol{y}_i$ can be difficult to obtain analytically in most situations, however, new computational algorithms for stochastic simulation or numerical integration are helpful in finding numerical approximations for these expectations. In the rest of this article we will use the linearized form of this conditional expectation, however other implementation can be explored.

Using the joint distribution of the complete data, we can obtain the conditional distribution of the sufficient statistics, in fact

$$\boldsymbol{\beta}_i|\tilde{\boldsymbol{y}}_i \sim N\left(\boldsymbol{D}\tilde{\boldsymbol{Z}}_i'\boldsymbol{W}_i(\tilde{\boldsymbol{y}}_i - \tilde{\boldsymbol{X}}_i\boldsymbol{\beta}), \boldsymbol{D} - \boldsymbol{D}\tilde{\boldsymbol{Z}}_i'\boldsymbol{W}_i\tilde{\boldsymbol{Z}}_i\boldsymbol{D}\right)$$

and

$$\boldsymbol{\epsilon}_i|\tilde{\boldsymbol{y}}_i \sim N\left(\tilde{\boldsymbol{y}}_i - \tilde{\boldsymbol{X}}_i\boldsymbol{\beta} - \tilde{\boldsymbol{Z}}_i\boldsymbol{\beta}_i, \boldsymbol{R}_i - \boldsymbol{R}_i\boldsymbol{W}_i\boldsymbol{R}_i\right).$$

Based on these results, we can find the first two moments of the conditional distribution of $\boldsymbol{\beta}_i$ given the observed data as

$$\tilde{\boldsymbol{\beta}}_i = E\left\{\boldsymbol{\beta}_i|\tilde{\boldsymbol{y}}_i, \boldsymbol{\beta}^{(\nu)}, \boldsymbol{D}^{(\nu)}, \boldsymbol{\Sigma}^{(\nu)}\right\} = \boldsymbol{D}^{(\nu)}\tilde{\boldsymbol{Z}}_i'\boldsymbol{W}_i^{(\nu)}(\tilde{\boldsymbol{y}}_i - \tilde{\boldsymbol{X}}_i\boldsymbol{\beta}^{(\nu)})$$

and

$$E\left\{\boldsymbol{\beta}_i\boldsymbol{\beta}_i'|\tilde{\boldsymbol{y}}_i, \boldsymbol{\beta}^{(\nu)}, \boldsymbol{D}^{(\nu)}, \boldsymbol{\Sigma}^{(\nu)}\right\} = \tilde{\boldsymbol{\beta}}_i\tilde{\boldsymbol{\beta}}_i' + \boldsymbol{D}^{(\nu)} - \boldsymbol{D}^{(\nu)}\tilde{\boldsymbol{Z}}_i'\boldsymbol{W}_i^{(\nu)}\tilde{\boldsymbol{Z}}_i\boldsymbol{D}^{(\nu)}.$$

Also we can find the first two moments of the conditional distribution of $\boldsymbol{\epsilon}_i$ given the observed data as

$$\tilde{\boldsymbol{\epsilon}}_i = E\left\{\boldsymbol{\epsilon}_i|\tilde{\boldsymbol{y}}_i, \boldsymbol{\beta}^{(\nu)}, \boldsymbol{D}^{(\nu)}, \boldsymbol{\Sigma}^{(\nu)}\right\} = \tilde{\boldsymbol{y}}_i - \tilde{\boldsymbol{X}}_i\boldsymbol{\beta}^{(\nu)} - \tilde{\boldsymbol{Z}}_i\tilde{\boldsymbol{\beta}}_i$$

and

$$E\left\{\boldsymbol{\epsilon}_i\boldsymbol{\epsilon}_i'|\tilde{\boldsymbol{y}}_i, \boldsymbol{\beta}^{(\nu)}, \boldsymbol{D}^{(\nu)}, \boldsymbol{\Sigma}^{(\nu)}\right\} = \tilde{\boldsymbol{\epsilon}}_i\tilde{\boldsymbol{\epsilon}}_i' + \boldsymbol{R}_i^{(\nu)} - \boldsymbol{R}_i^{(\nu)}\boldsymbol{W}_i^{(\nu)}\boldsymbol{R}_i^{(\nu)}.$$

However from the analytical point of view it seems easier to use the conditional expectation of the rows of the $\boldsymbol{E}_i$ matrix, that is $\boldsymbol{E}_{i[j]}$. If $\tilde{\boldsymbol{E}}_{i[j]}' = E\left\{\boldsymbol{E}_{i[j]}'|\tilde{\boldsymbol{y}}_i, \boldsymbol{\beta}^{(\nu)}, \boldsymbol{D}^{(\nu)}, \boldsymbol{\Sigma}^{(\nu)}\right\}$ and

$$E\left\{\boldsymbol{E}_{i[j]}\boldsymbol{E}_{i[j]}'|\tilde{\boldsymbol{y}}_i, \boldsymbol{\beta}^{(\nu)}, \boldsymbol{D}^{(\nu)}, \boldsymbol{\Sigma}^{(\nu)}\right\} = \tilde{\boldsymbol{E}}_{i[j]}\tilde{\boldsymbol{E}}_{i[j]}' + \boldsymbol{\Sigma}^{(\nu)} - \boldsymbol{\Sigma}^{(\nu)}\boldsymbol{W}_{i[j,j]}^{(\nu)}\boldsymbol{\Sigma}^{(\nu)}$$

where $\boldsymbol{W}_{i[j,j]}^{(\nu)}$ is a $p \times p$ matrix with the elements of $\boldsymbol{W}_i^{(\nu)}$ corresponding to the observation taken at time j. To orient the reader the element $(jk, j'k')$ of the $\tilde{\boldsymbol{V}}_i = \boldsymbol{W}_i^{-1}$ is

$$\tilde{\boldsymbol{V}}_{i[jk,j'k']} = \sum_{l=1}^{p}\sum_{m=1}^{p} \tilde{z}_{ijkl}\boldsymbol{D}_{lm}\tilde{z}_{ij'k'm} + \sigma_{kk'}^2 I(j = j')$$

where $\sigma_{kk'}^2$ is the element (k, k') of $\boldsymbol{\Sigma}$ and $\boldsymbol{D}_{lm}$ is the element (l, m) of $\boldsymbol{D}$.

The M-step is

$$\boldsymbol{D}^{(\nu+1)} = \frac{1}{N}\sum_{i=1}^{N}\left(\tilde{\boldsymbol{\beta}}_i\tilde{\boldsymbol{\beta}}_i' + \boldsymbol{D}^{(\nu)} - \boldsymbol{D}^{(\nu)}\tilde{\boldsymbol{Z}}_i'\boldsymbol{W}_i^{(\nu)}\tilde{\boldsymbol{Z}}_i\boldsymbol{D}^{(\nu)}\right)$$

and

$$\boldsymbol{\Sigma}^{(\nu+1)} = \frac{1}{\sum_{i=1}^{N} n_i}\sum_{i=1}^{N}\sum_{j=1}^{n_i}\left(\tilde{\boldsymbol{E}}_{i[j]}\tilde{\boldsymbol{E}}_{i[j]}' + \boldsymbol{\Sigma}^{(\nu)} - \boldsymbol{\Sigma}^{(\nu)}\boldsymbol{W}_{i[j,j]}^{(\nu)}\boldsymbol{\Sigma}^{(\nu)}\right)$$

or equivalently

$$\boldsymbol{\Sigma}^{(\nu+1)} = \frac{1}{\sum_{i=1}^{N} n_i}\sum_{i=1}^{N}\sum_{j=1}^{n_i}\left(\tilde{\boldsymbol{E}}_{i[j]}\tilde{\boldsymbol{E}}_{i[j]}' + \boldsymbol{\Sigma}^{(\nu)} - \boldsymbol{\Sigma}^{(\nu)}H_{ij}\boldsymbol{W}_i^{(\nu)}H_{ij}'\boldsymbol{\Sigma}^{(\nu)}\right)$$

where $H_{ij} = a_{ij} \otimes \boldsymbol{I}_p$ and a_{ij} is the jth row of the identity matrix $\boldsymbol{I}_i$.

3 Estimation with Missing Data

In order to consider the unbalanced case where the measurements Are made at different times or there are missing data, we first introduce an indicator matrix $\boldsymbol{O}_i$ with only *ones* and *zeros*, which is generated from the identity matrix by deleting the rows corresponding to the missing observations. In this case, the observed multi-response vector $\boldsymbol{y}_i^o$ can be written as a function of the balanced multi-response vector $\boldsymbol{y}$ as $\boldsymbol{y}_i^o = \boldsymbol{O}_i\boldsymbol{y}_i$. The model is pre-multiplied by this matrix and can be written as

$$\tilde{\boldsymbol{y}}_i^o = \tilde{\boldsymbol{X}}_i^o\boldsymbol{\beta} + \tilde{\boldsymbol{Z}}_i^o\boldsymbol{\beta}_i + \boldsymbol{\epsilon}_i^o$$

where $\tilde{\boldsymbol{X}}_i^o = \boldsymbol{O}_i\tilde{\boldsymbol{X}}_i$, $\tilde{\boldsymbol{Z}}_i^o = \boldsymbol{O}_i\tilde{\boldsymbol{Z}}_i$, and $\boldsymbol{\epsilon}_i^o = \boldsymbol{O}_i\boldsymbol{\epsilon}_i$. Computationally, we do not need to pre-multiply by $\boldsymbol{O}_i$ since this is equivalent to including in $\tilde{\boldsymbol{y}}_i$, $\tilde{\boldsymbol{X}}_i$, and $\tilde{\boldsymbol{Z}}_i$ only the observed responses and their corresponding covariates.

Similar to the complete data, the E-step starts with

$$\boldsymbol{\beta}^{(\nu)} = \left(\sum_{i=1}^{N} \tilde{\boldsymbol{X}}_i^{o\prime}\boldsymbol{W}_i^{o(\nu)}\tilde{\boldsymbol{X}}_i^o\right)^{-1} \sum_{i=1}^{N} \tilde{\boldsymbol{X}}_i^{o\prime}\boldsymbol{W}_i^{o(\nu)}\tilde{\boldsymbol{y}}_i^o$$

where $\boldsymbol{W}_i^{o(\nu)} = \left(\tilde{\boldsymbol{Z}}_i^o\boldsymbol{D}^{(\nu)}\tilde{\boldsymbol{Z}}_i^{o\prime} + \boldsymbol{O}_i\boldsymbol{R}_i^{(\nu)}\boldsymbol{O}_i'\right)^{-1}$. The covariance matrix is $\boldsymbol{V}_i^o = \boldsymbol{O}_i\boldsymbol{V}_i\boldsymbol{O}_i^{\mathrm{T}}$ (i.e. $\boldsymbol{y}_i^o \sim \mathrm{MVN}(\boldsymbol{X}_i^o\boldsymbol{\beta}, \boldsymbol{V}_i^o)$, where $\boldsymbol{X}_i^o = \boldsymbol{O}_i\boldsymbol{X}_i$).

Using the joint distribution of the complete data, we can obtain the conditional distribution of the sufficient statistics, in fact

$$\boldsymbol{\beta}_i|\tilde{\boldsymbol{y}}_i^o \sim N\left(\boldsymbol{D}\tilde{\boldsymbol{Z}}_i^{o\prime}\boldsymbol{W}_i^o(\tilde{\boldsymbol{y}}_i^o - \tilde{\boldsymbol{X}}_i^o\boldsymbol{\beta}), \boldsymbol{D} - \boldsymbol{D}\tilde{\boldsymbol{Z}}_i^{o\prime}\boldsymbol{W}_i^o\tilde{\boldsymbol{Z}}_i^o\boldsymbol{D}\right)$$

and

$$\boldsymbol{\epsilon}_i|\tilde{\boldsymbol{y}}_i^o \sim N\left(\boldsymbol{R}_i\boldsymbol{O}_i'\boldsymbol{W}_i^o(\tilde{\boldsymbol{y}}_i^o - \tilde{\boldsymbol{X}}_i^o\boldsymbol{\beta}), \boldsymbol{R}_i - \boldsymbol{R}_i\boldsymbol{O}_i'\boldsymbol{W}_i^o\boldsymbol{O}_i\boldsymbol{R}_i\right).$$

Based on these results, we can find the first two moments of the conditional distribution of $\boldsymbol{\beta}_i$ given the observed data as

$$\tilde{\boldsymbol{\beta}}_i = E\left\{\boldsymbol{\beta}_i|\widetilde{\boldsymbol{y}_i^o}, \boldsymbol{\beta}^{(\nu)}, \boldsymbol{D}^{(\nu)}, \boldsymbol{\Sigma}^{(\nu)}\right\} = \boldsymbol{D}^{(\nu)}\tilde{\boldsymbol{Z}}_i^{o\prime}\boldsymbol{W}_i^{o(\nu)}\left(\tilde{\boldsymbol{y}}_i^o - \tilde{\boldsymbol{X}}_i^o\boldsymbol{\beta}^{(\nu)}\right)$$

and

$$E\left\{\boldsymbol{\beta}_i\boldsymbol{\beta}_i'|\widetilde{\boldsymbol{y}_i^o}, \boldsymbol{\beta}^{(\nu)}, \boldsymbol{D}^{(\nu)}, \boldsymbol{\Sigma}^{(\nu)}\right\} = \tilde{\boldsymbol{\beta}}_i\tilde{\boldsymbol{\beta}}_i' + \boldsymbol{D}^{(\nu)} - \boldsymbol{D}^{(\nu)}\tilde{\boldsymbol{Z}}_i^{o\prime}\boldsymbol{W}_i^{(\nu)}\tilde{\boldsymbol{Z}}_i^o\boldsymbol{D}^{(\nu)}.$$

We can also find the first two moments of the conditional distribution of $\boldsymbol{\epsilon}_i$ given the observed data as

$$\tilde{\boldsymbol{\epsilon}}_i = E\left\{\boldsymbol{\epsilon}_i|\widetilde{\boldsymbol{y}_i^o}, \boldsymbol{\beta}^{(\nu)}, \boldsymbol{D}^{(\nu)}, \boldsymbol{\Sigma}^{(\nu)}\right\} = \boldsymbol{R}_i\boldsymbol{O}_i'\boldsymbol{W}_i^o(\tilde{\boldsymbol{y}}_i^o - \tilde{\boldsymbol{X}}_i^o\boldsymbol{\beta})$$

and

$$E\left\{\boldsymbol{\epsilon}_i\boldsymbol{\epsilon}_i'|\widetilde{\boldsymbol{y}_i^o}, \boldsymbol{\beta}^{(\nu)}, \boldsymbol{D}^{(\nu)}, \boldsymbol{\Sigma}^{(\nu)}\right\} = \tilde{\boldsymbol{\epsilon}}_i\tilde{\boldsymbol{\epsilon}}_i' + \boldsymbol{R}_i - \boldsymbol{R}_i\boldsymbol{O}_i'\boldsymbol{W}_i^o\boldsymbol{O}_i\boldsymbol{R}_i.$$

Again, from the analytical point of view it is easier to consider the conditional expectation on the rows of the $\boldsymbol{E}_i$ matrix, this is $\boldsymbol{E}_{i[j]}$. If $\tilde{\boldsymbol{E}}_{i[j]}' = E\left\{\boldsymbol{E}_{i[j]}'|\widetilde{\boldsymbol{y}_i^o}, \boldsymbol{\beta}^{(\nu)}, \boldsymbol{D}^{(\nu)}, \boldsymbol{\Sigma}^{(\nu)}\right\}$ and

$$E\left\{\boldsymbol{E}_{i[j]}\boldsymbol{E}_{i[j]}'|\tilde{\boldsymbol{y}}_i, \boldsymbol{\beta}^{(\nu)}, \boldsymbol{D}^{(\nu)}, \boldsymbol{\Sigma}^{(\nu)}\right\} = \tilde{\boldsymbol{E}}_{i[j]}\tilde{\boldsymbol{E}}_{i[j]}' + \boldsymbol{\Sigma}^{(\nu)} - \boldsymbol{\Sigma}^{(\nu)}H_{ij}\boldsymbol{O}_i'\boldsymbol{W}_i^o\boldsymbol{O}_iH_{ij}'\boldsymbol{\Sigma}^{(\nu)}$$

where H_{ij} is as defined before.

The M-step is then

$$\boldsymbol{D}^{(\nu+1)} = \frac{1}{N}\sum_{i=1}^{N}\left(\tilde{\boldsymbol{\beta}}_i\tilde{\boldsymbol{\beta}}_i' + \boldsymbol{D}^{(\nu)} - \boldsymbol{D}^{(\nu)}\tilde{\boldsymbol{Z}}_i^{o\prime}\boldsymbol{W}_i^{o(\nu)}\tilde{\boldsymbol{Z}}_i^{o}\boldsymbol{D}^{(\nu)}\right)$$

and

$$\boldsymbol{\Sigma}^{(\nu+1)} = \frac{1}{\sum_{i=1}^{N} n_i}\sum_{i=1}^{N}\sum_{j=1}^{n_i}\left(\tilde{\boldsymbol{E}}_{i[j]}\tilde{\boldsymbol{E}}_{i[j]}' + \boldsymbol{\Sigma}^{(\nu)} - \Sigma^{(\nu)}H_{ij}\boldsymbol{O}_i'\boldsymbol{W}_i^{o(\nu)}\boldsymbol{O}_iH_{ij}'\boldsymbol{\Sigma}^{(\nu)}\right).$$

4 An Example

Model and Data

It is well known in obstetrics that serum beta-subunit human chorionic gonadotropin (β-HCG) and estardiol experience dramatic changes in women during the first trimester of pregnancy. It has been also established that values of the β-HCG are different in women who have normal pregnancies with terminal deliveries than in women who have spontaneous abortions or other types of adverse pregnancy outcomes.

In a follow-up study of 161 young women, representing 161 different pregnancies over a period of 2 years in a private obstetrics clinic in Santiago, Chile, values of the β-HCG and of Estradiol were measured during the first 80 days of gestational age. The women were classified as normal, if they had a normal delivery, or as abnormal if they had any complication resulting in a non-terminal delivery and loss of the fetus. Missingness rates for the two responses are shown in Table 1. The differences among the two groups could be explain by the presence of informative censoring. Further research in this direction is undergoing and will be included in a future communication.

Table 1. Missingness rates (%) by response

	Normal Pregnancies	Abnormal Pregnancies
β-HCG	3	0
Estradiol	27	58

Mean values of β-HCG by gestational age (days) show a non-linear relationship, with a threshold after 50 days of pregnancy. The logistic curve shown in Figure 1 is a reasonable function to describe the changes in β-HCG in the log scale over time. Letting log β-HCG be response $k = 1$, the proposed model for the ith women at time j in group ℓ, where $\ell = 1$ for the normal pregnancy and $\ell = 2$ for the abnormal pregnancy group, is

$$y_{ij1}^{\ell} = \frac{\beta_{\ell 1} + b_{i1\ell}}{1 + \exp\{(\beta_{\ell 2} - t_{ij})/\beta_{\ell 3}\}} + e_{ij1}^{\ell}. \tag{1}$$

A four parameter logistic curve was fit to the data, however, the extra additive parameter was found no significant.

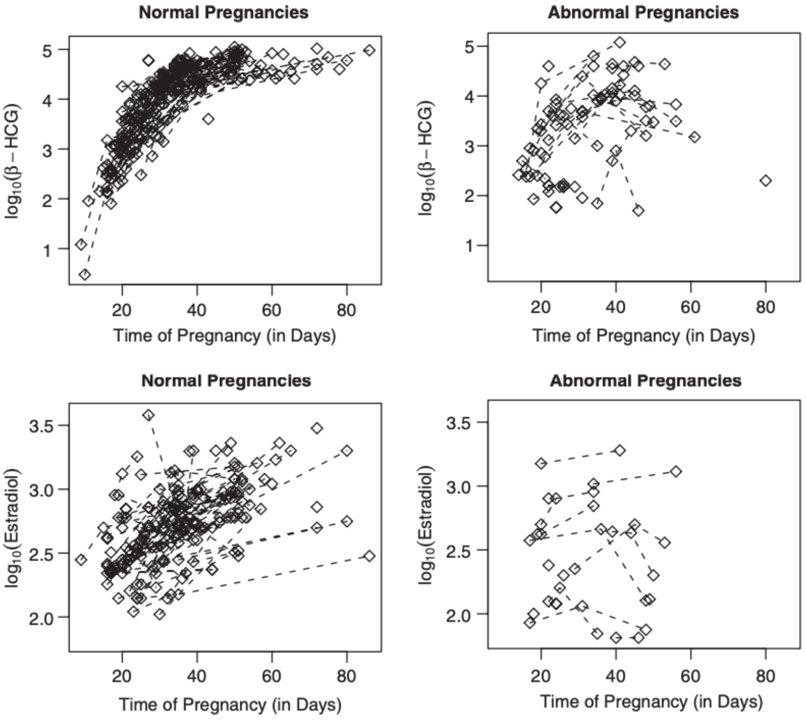


Fig. 1. Logistic and linear curve for the β-HCG and the Estradiol.

Mean values of the log estradiol by gestational age days show a linear relationship. The linear model shown Figure 1 is a reasonable function to describe the changes in the estradiol in the log scale across the days of pregnancy. Letting Estradiol be response $k = 2$, the proposed model for the ith subject at time j in group ℓ is

$$y_{ij2}^{\ell} = \beta_{\ell 4} + \beta_{\ell 5} t_{ij} + b_{i2\ell} + e_{ij2}^{\ell}. \tag{2}$$

The random components for this models are assumed to follow a normal distribution, that is, $b_{i\ell} \sim \text{MVN}(0, \boldsymbol{D}_\ell)$ and $e_{ij}^{\ell} \sim \text{MVN}(0, \boldsymbol{R}_i)$, with $\boldsymbol{R}_i = \boldsymbol{\Sigma}_\ell \otimes \mathbf{I}_i$, where the e_{ij}^{ℓ}'s are assumed to be independent of the $b_{i\ell}$'s. Only one random effect was included in our models mainly due to the fact that 61 per cent of the women have two or fewer observations, and 94 per cent have three or fewer observations. Figure 2 shows time profiles for selected normal and abnormal subjects. The parameters $(\beta_{\ell 1}, \beta_{\ell 2}, \beta_{\ell 3})$ and $(\beta_{\ell 4}, \beta_{\ell 5})$ represent the population parameters of the logistic curve and of the linear model, respectively.

If we let $f_1^{\ell}(\boldsymbol{\beta}_\ell, t_{ij}) = (\beta_{\ell 1} + b_{i1\ell})/(1 + \exp\{(\beta_{\ell 2} - t_{ij})/\beta_{\ell 3}\})$, and $f_2^{\ell}(\boldsymbol{\beta}_\ell, t_{ij}) = \beta_{\ell 4} + \beta_{\ell 5} t_{ij} + b_{i2\ell}$ then the linearized version of the model for the ith subject at jth time in group ℓ is

$$\tilde{\boldsymbol{y}}_{ij}^{\ell} = \tilde{\boldsymbol{X}}_{ij}^{\ell} \boldsymbol{\beta}_\ell + \tilde{\boldsymbol{Z}}_{ij}^{\ell} \boldsymbol{\beta}_{i\ell} + \boldsymbol{\epsilon}_{ij}^{\ell} \tag{3}$$

where the working response variable is a 2 × 1 vector with elements

$$\tilde{\boldsymbol{y}}_{ijk}^{\ell} = \boldsymbol{y}_{ijk}^{\ell} - f_k^{\ell}(\boldsymbol{\beta}_\ell, t_{ij}) + \tilde{\boldsymbol{x}}_{ijk}^{\ell} \boldsymbol{\beta}_\ell + \tilde{\boldsymbol{z}}_{ijk}^{\ell} \boldsymbol{\beta}_{i\ell} \tag{4}$$

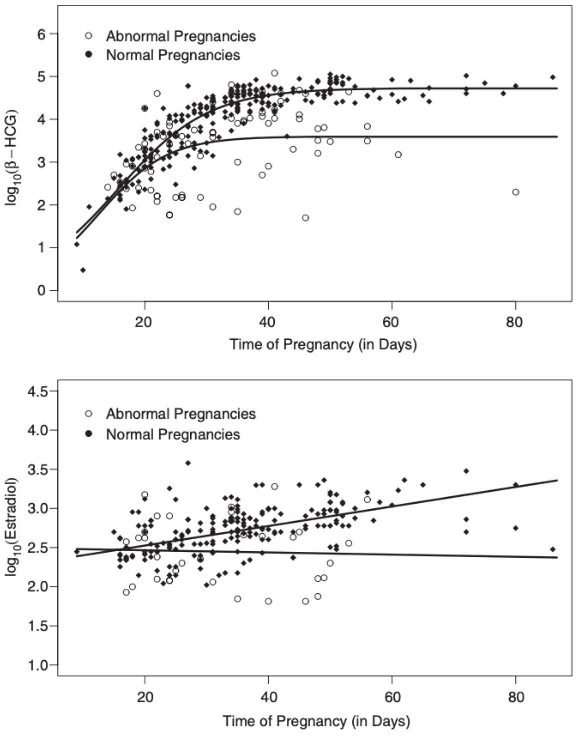


Fig. 2. Time profiles for randomly selected normal and abnormal subjects.

where $\boldsymbol{\beta}^{'}_{\ell} = (\beta_{\ell 1}, \beta_{\ell 2}, \beta_{\ell 3}, \beta_{\ell 4}, \beta_{\ell 5})$, $\boldsymbol{\beta}'_{i\ell} = (b_{i1\ell}, b_{i2\ell})$, the working design matrix $\tilde{\boldsymbol{X}}^{\ell}_{ij}$ of dimension 2×5 has elements

$$\tilde{\boldsymbol{X}}'^{\ell}_{ij} = \begin{pmatrix} w^{\ell}_{ij} & 0 \\ -f^{\ell}_1(\boldsymbol{\beta}_\ell, t_{ij})\beta^{-1}_{\ell 3} \exp\{(\beta_{\ell 2} - t_{ij})/\beta_{\ell 3}\} w^{\ell}_{ij} & 0 \\ f^{\ell}_1(\boldsymbol{\beta}_\ell, t_{ij}) \exp\{(\beta_{\ell 2} - t_{ij})/\beta_{\ell 3}\}(\beta_{\ell 2} - t_{ij})/\beta^2_{\ell 3} w^{\ell}_{ij} & 0 \\ 0 & 1 \\ 0 & t_{ij} \end{pmatrix} \tag{5}$$

where $w^{\ell}_{ij} = 1/(1 + \exp\{(\beta_{\ell 2} - t_{ij})/\beta_{\ell 3}\})$, and the working design matrix $\tilde{\boldsymbol{Z}}^{\ell}_{ij}$ of dimension 2×2 has elements

$$\tilde{\boldsymbol{Z}}^{\ell}_{ij} = \begin{pmatrix} w^{\ell}_{ij} & 0 \\ 0 & 1 \end{pmatrix} \tag{6}$$

which depend on the values of the unknown population parameters $\beta_{\ell 2}$ and $\beta_{\ell 3}$. The covariance matrix of the $\boldsymbol{\epsilon}^{\ell}_{ij}$ is $\boldsymbol{\Sigma}_\ell$ of dimension 2×2. The covariance matrix of the random effects, $\boldsymbol{D}_\ell$, is a 2×2 matrix.

To illustrate, the 13th subject of this study has measurements at time $t_{13,1} = 28$, $t_{13,2} = 33$, and $t_{13,3} = 41$ which in matrix notation can be represented as

$$\boldsymbol{Y}_{13} = \begin{pmatrix} 3.32 & \cdot \\ \cdot & 2.58 \\ 4.20 & 2.70 \end{pmatrix} \tag{7}$$

where the $\cdot$ represent missing data. The corresponding vector format of the responses is $\boldsymbol{y}'_{13} = (3.32, \cdot, 4.20, \cdot, 2.58, 2.70)$. Because of the missing data observed in this subject, we introduce the $\boldsymbol{O}_{13}$ matrix as

$$\boldsymbol{O}_{13} = \begin{pmatrix} 1\,0\,0\,0\,0\,0 \\ 0\,0\,1\,0\,0\,0 \\ 0\,0\,0\,0\,1\,0 \\ 0\,0\,0\,0\,0\,1 \end{pmatrix} \tag{8}$$

to eliminate the missing data element in $\boldsymbol{y}_{13}$. Therefore, the observed response vector $\boldsymbol{y}^o_{13} = \boldsymbol{O}_{13}\boldsymbol{y}_{13}$ has elements $\boldsymbol{y}^{o\prime}_{13} = (3.32, 4.20, 2.58, 2.70)$. The covariance matrix of the error term for the complete data is

$$\boldsymbol{R}_{13} = \begin{pmatrix} \sigma_{11} & 0 & 0 & \sigma_{12} & 0 & 0 \\ 0 & \sigma_{11} & 0 & 0 & \sigma_{12} & 0 \\ 0 & 0 & \sigma_{11} & 0 & 0 & \sigma_{12} \\ \sigma_{21} & 0 & 0 & \sigma_{22} & 0 & 0 \\ 0 & \sigma_{21} & 0 & 0 & \sigma_{22} & 0 \\ 0 & 0 & \sigma_{21} & 0 & 0 & \sigma_{22} \end{pmatrix}.$$

However, the covariance matrix of the error term for the observed data is

$$\boldsymbol{O}_{13}\boldsymbol{R}_{13}\boldsymbol{O}'_{13} = \begin{pmatrix} \sigma_{11} & 0 & 0 & 0 \\ 0 & \sigma_{11} & 0 & \sigma_{12} \\ 0 & 0 & \sigma_{22} & 0 \\ 0 & \sigma_{21} & 0 & \sigma_{22} \end{pmatrix}.$$

The H_{ij} matrices in the M-step for this subject are $H_{13,1} = (1, 0, 0, 0)$, $H_{13,2} = (0, 0, 1, 0)$ and

$$H_{13,3} = \begin{pmatrix} 0\,1\,0\,0 \\ 0\,0\,0\,1 \end{pmatrix}.$$

Results

We implemented the algorithm described in the previous sections as a function in S-Plus Version 6 to fit the models. We used the NLME software of Pinheiro and Bates (22) to fit the models (1) and (2) to obtain initial estimates of the fixed effects, $(\beta_{\ell 1}, \beta_{\ell 2}, \beta_{\ell 3})$ and $(\beta_{\ell 4}, \beta_{\ell 5})$, random effects, $b_{i1\ell}$ and $b_{i2\ell}$, and variance components estimates. The S-Plus code for this example can be obtained by request from the first author.

When we ran the program we achieved convergence within a maximum of 27 iterations. Table 2 gives the ML fixed effects estimates.

The elements of $\hat{\boldsymbol{\Sigma}}_\ell$, representing the estimates of the within-subject variability, are

$$\hat{\boldsymbol{\Sigma}}_1 = \begin{bmatrix} 0.0926 & 0.0024 \\ 0.0024 & 0.0277 \end{bmatrix}, \qquad \hat{\boldsymbol{\Sigma}}_2 = \begin{bmatrix} 0.1961 & 0.0310 \\ 0.0310 & 0.0274 \end{bmatrix}.$$

Table 2. Fixed Effects Estimates

	Normal Pregnancies	Abnormal Pregnancies
Parameter	($\ell = 1$)	($\ell = 2$)
$\beta_{\ell 1}$	4.7281	3.5984
$\beta_{\ell 2}$	15.6006	12.5977
$\beta_{\ell 3}$	7.2751	5.4609
$\beta_{\ell 4}$	2.2717	2.4946
$\beta_{\ell 5}$	0.0129	-0.0015

The elements of $\hat{\boldsymbol{D}}_\ell$, representing the estimates of the variances of the random effects, are

$$\hat{\boldsymbol{D}}_1 = \begin{bmatrix} 0.0310 & 0.0126 \\ 0.0126 & 0.0404 \end{bmatrix}, \qquad \hat{\boldsymbol{D}}_2 = \begin{bmatrix} 0.6338 & 0.1969 \\ 0.1969 & 0.1310 \end{bmatrix}.$$

Observed differences in the group-specific estimates of the variance components suggest that there is significantly more between-subject variability in the abnormal group than in the normal group as is revealed in Figure 2.

In order to test the differences between the two groups, we compare four alternative models. The first model (Model 1) assume no differences among the two groups, and common curve for the two responses was fit ($\boldsymbol{\beta}_1 = \boldsymbol{\beta}_2, \boldsymbol{D}_1 = \boldsymbol{D}_2, \boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$). The second model (Model 2) assume a common curve for Estradiol but different curve for β-HCG. The third model (Model 3) assume a common curve for β-HCG and a group specific curve for Estradiol. The fourth model (Model 4), previously described, considered different curves for both responses, β-HCG and Estradiol.

As shown in Table 3 the best model among these four considered is the model that have different curve for the two groups and both responses. If the purpose of the analysis is to test differences among the two groups, the multiple response model provides more power to detect statistical differences.

Table 3. Summary of model fitting

	Number of			
Model	Parameters	$-2 \log L$	AIC	BIC
1	11	490.6	512.6	546.5
2	15	482.3	512.3	558.6
3	16	347.9	379.6	429.2
4	22	310.0	354.1	421.8

Implementing the algorithm in SAS NLMIXED

For the example presented above, we also implemented a variant of our algorithm using SAS NLMIXED with the first-order approximation method of Beal and Sheiner (23), which instead of evaluating the current values of the random effects $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}$ at each iteration, evaluates instead at $\boldsymbol{\beta} = 0$. It should be noted that this approach leads to suboptimal parameter estimates (13). Because the random effects are included linearly in our example this option produces the same results to ours.

To implement the algorithm for the data in our example using SAS NLMIXED we used the fact that the joint distribution of the $\boldsymbol{y}_{ij}$ can be written as the product of the conditional distribution of y_{ij2} given y_{ij1} and the marginal distribution of y_{ij1}. For missing observations on one or the other response the contribution to the likelihood comes from the marginal density of the observed response. This effectively tricks NLMIXED into treating the bivariate model as a univariate one.

The SAS code for modelling the responses in the normal pregnancy outcome group is included in the Appendix.

5 Discussion

We have extended the nonlinear random effects model for a single response to the case of multiple responses, allowing for arbitrary patterns of observed and missing data. We used the EM algorithm to obtain parameter estimates. Finally we illustrated the approach with an example using data coming from a study involving bivariate hormone measurements taken during the first trimester of pregnancy.

The main advantage of the multiple response model lies in its ability to utilize the inherent covariance structure for a truly multivariate response. This results in more efficient estimation of the parameters. In the case of a single response, the model reduces to the Lindstrom and Bates model (13).

The example we have used to illustrate the multiple response model may contain informative censoring, since the abnormal group have significantly more missing data in Estradiol than the normal pregnancy group. Further research is ongoing to extend this model to contemplate informative censoring.

Bibliography

[1] Grizzle JE, Allen DM. Analysis of growth and dose-response curves. *Biometrics* 1969; **25**:357–381.

[2] Khatri CG. A note on a manova model applied to problems in growth curve. *Annals of the Institute of Statistical Mathematics* 1966; **8**:75–86.

[3] Potthoff R, Roy SN. A generalized multivariate analysis of variance model useful especially for growth curve problems. *Biometrika* 1964; **51**:313–326.

[4] Rao CR. The theory of least squares when the parameters are stochastic and its application to the analysis of growth curves. *Biometrika* 1965; **52**:447–468.

[5] Schafer JL, Yucel RM. Computational strategies for multivariate linear mixed-effects models with missing values. *Journal of Computational and Graphical Statistics* 2002; **11**:437–457.

[6] Laird NM, Ware JH. Random-effects models for longitudinal data. *Biometrics* 1982; **38**:963–974.

[7] Reinsel G. Estimation and prediction in a multivariate random-effects generalized linear model. *Journal of the American Statistical Association* 1984; **79**:406–414.

[8] Mickey RM, Shema SJ, Vacek, PM, Bell DY. Analysis of multiple outcome variables measured longitudinally. *Computational Statistics and Data Analysis* 1994; **17**:17–33.

[9] Zucker DM, Zerbe GO, Wu MC. Inference for the Association Between Coefficients in a Multivariate Growth Curve Model. *Biometrics* 1995 **51**:413–424.

[10] Shah A, Laird N, Schoenfeld D. A random-effects model for multiple characteristics with possibly missing data. *Journal of American Statistical Association* 1997; **92**:775–779.

[11] Beal SL, Sheiner LB. *NONMEM users' guide, Part I.* San Francisco: Division of Clinical Pharmacology, University of California. 1979.

[12] Steimer JL, Mallet A, Golmard JL, Boisvieux JF. Alternative approaches to estimation of population pharmacokinetic parameters: Comparison with the non-linear mixed effects model. *Drug Metabolism Reviews* 1984; **15**:265–292.

[13] Lindstrom MJ, Bates DM. Nonlinear random effects models for repeated measures data. *Biometrics* 1990; **46**:673–687.

[14] Hirst K, Zerbe GO, Boyle DW, Wilkening RB. On nonlinear random effects models for repeated measurements. *Communications in Statistics B, Simulation and Computation* 1991; **20**:463–478.

[15] Beal SL, Sheiner LB. *NONMEM users' guide, Part VII, Conditional Estimation Methods.* San Francisco: NONMEM Project Group, University of California. 1992.

[16] Vonesh EF, Carter RL. Mixed-effects nonlinear regression for unbalanced repeated measures. *Biometrics* 1992; **48**:1–18.

[17] Mentre F, Gomeni R. A two step iterative algorithm for estimation in nonlinear mixed effects models with an evaluation in population pharmacokinetics. *Journal of Biopharmaceutical Statistics* 1995; **5**:141–158.

[18] Wolfinger R. Laplace's approximation for nonlinear models. *Biometrika* 1993; **80**:791–795.

[19] Pinheiro JC, Bates DM. Approximations to the log-likelihood function in the nonlinear mixed effects models. *Journal of Computational and Graphical Statistics* 1995; **4(1)**:12–35.

[20] Davidian M, Giltinan, DM. *Nonlinear Models for Repeated Measurement Data.* Chapman & Hall, 1995.

[21] Walker S. An EM algorithm for nonlinear random effects model. *Biometrics* 1996; **52**:934–944.

[22] Pinheiro JC, Bates DM. *Mixed-effects models in S and S-PLUS.* New York: Springer, 2000.

[23] Beal SL, Sheiner LB. Heteroskedastic Nonlinear Regression *Technometrics* 1988. **30**:327–338

SAS NLMIXED code for the normal pregnancy outcome group in the example in section 4.

```
 data transform;
   set MYDATA;
   bHCG=log10(bHCG);
   E2=log10(E2);
 run;

data univar;
  set transform;
  Y=bHCG; Var=1; if bHCG=. then Var=3; if E2=. then Var=1; output;
  Y=E2;   Var=2; if bHCG=. then Var=3; if E2=. then Var=1; output;
  keep Subject Time bHCG E2 Var Y;

data univar;
  set univar;
  if Y=. then delete;

proc nlmixed data=univar method=firo;
  parms B1=5 B2=10 B3=5 B4=2 B5=0.5 S11=.1
        S22=.1 S12=0 D11=1 D22=1 D12=0;
  bounds  B5 > 0, S11 > 0, S22 > 0, D11 > 0, D22 > 0;
  * Marginal distribution of bHCG when both responses or only bHCG is observed;
  if Var=1 then do;
    M1=(B1+u1)/(1+exp((B2-Time)/B3));
    MZ=M1;
    VZ=S11;
  end;
  * Conditional distribution of E2 given bHCG when both responses are observed;
  else if Var=2 then do;
    M2=B4+B5*Time+u2;
    MZ=M2+S12*(bHCG-M1)/S11;
    VZ=S22-S12*S12/S11;
  end;
  * Marginal distribution of E2 when only E2 is observed;
  else do;
    M2=B4+B5*Time+u2;
    MZ=M2;
    VZ=S22;
  end;
  model Y ~ normal(MZ,VZ);
  random u1 u2 ~ normal([0,0],[D11, D12, D22]) subject=subject;
run;
```

Article 2.3

Discriminant Analysis for Longitudinal Data with Multiple Continuous Responses and Possibly Missing Data

Guillermo Marshall, Rolando de la Cruz-Mesía,
Fernando A. Quintana and Anna E. Barón

Pontificia Universidad Católica de Chile and
University of Colorado HSC, Denver, U.S.A.

Abstract. Multiple outcomes are often used to properly characterize an effect of interest. This paper discusses model-based statistical methods for the classification of units into one of two or more groups where, for each unit, repeated measurements over time are obtained on each outcome. We relate the observed outcomes using multivariate nonlinear mixed-effects models to describe evolutions in different groups. Due to its flexibility, the random-effects approach for the joint modeling of multiple outcomes can be used to estimate population parameters for a discriminant model that classifies units into distinct predefined groups or populations. Parameter estimation is done via the EM algorithm with a linear approximation step. We conduct a simulation study that sheds light on how the linear approximation affects the classification results. An example is presented using data from a study in 161 pregnant women in Santiago, Chile in order to predict normal versus abnormal pregnancy outcomes. [1]

1 Introduction

Multivariate longitudinal data arise when a set of different responses on the same individual are measured repeatedly over time. Monitoring the extent or severity of disease over time using several clinical parameters is a common practice in the process of medical decision making. For instance patients meeting some threshold level of severity may be offered new or more aggressive treatments. Our motivating example concerns pregnant women. To assess risk factors for and to detect a number of complications during pregnancy, a variety of quantities or characteristics are measured at the prenatal examinations.

Some hormones produced during pregnancy are present in much larger amounts than in the non-pregnant state. In contrast, other hormones are unique in the sense of being present only during pregnancy. The endocrine changes that a pregnant women undergoes are usually related to pregnancy maintenance and mainly for the benefit of the fetus, whose metabolic needs vary greatly during gestation. The maternal endocrine and metabolic environment must adapt to these varying fetal requirements.

First trimester spontaneous abortions are common in both unassisted pregnancies and pregnancies that result from treatment with reproduction enhancing techniques. A reliable and inexpensive diagnostic test to differentiate between viable pregnancies and pregnancies with eventual early adverse outcome might reduce the psychological tension and anxiety present in many of

[1] Marshall G, De la Cruz-Mesía, R, Quintana, FA and Barón A E. (2009) Discriminant Analysis for Multivariate Longitudinal Markers with Possibly Missing Data. *Biometrics*, 65, 69-80

these patients, and also reduce the cost by making the treatment more effective. On the other hand, a more careful follow up might reduce the risks associated with abnormal pregnancies for patients who are in a higher risk group according to such a test.

Ultrasound examination is effective for evaluation of ongoing pregnancies, but a gestational sac is not reliably visible until 33–37 days after the luteinizing hormone surge (Shapiro et al., 1992). As a result of this inability of US to identify very early pregnancy abnormalities, there is an ongoing effort to institute a method that can forecast pregnancy outcome. Various studies have investigated hormones like estradiol, beta-subunit human chorionic gonadotropin (β-HCG), among others, and their relationship to pregnancy outcome after in-vitro fertilization (see e.g. Yamashita et al., 1989).

For continuous longitudinal data, when only a single outcome is observed, extensions of classical discriminant analysis to longitudinal data have been considered. Verbeke and Lesaffre (1996) proposed a linear mixed-effects model with random-effects sampled from a mixture of normal distributions. Verbeke and Molenberghs (2000) indicated that the classification rule implied by Verbeke and Lesaffre's model is equivalent to the discriminant function proposed by Tomasko, Helms and Snapinn (1999). Further developments along this direction have been discussed in Brant et al., (2003) and Wernecke et al (2004). De la Cruz and Quintana (2006) and Marshall and Baron (2000) using Bayesian and classical approaches, respectively, developed a mixed effects model for classification of hormone trajectories into pre-defined groups. Brown, Kenward, and Bassett (2000), with the goal of identifying Olympic athletes who use growth hormone injections, developed a Bayesian method that defines trajectory classes based on a training dataset with known classification. Recently, De la Cruz, Quintana, Muller (2005) developed a semiparametric Bayesian approach for classification of longitudinal markers. They defined a suitable extension of hierarchical models to allow such classification.

Some work has been done on longitudinal data with multiple outcomes using multivariate nonlinear mixed-effects models (M-NLMMs). Fitting separate models to each outcome is unsatisfactory because no correlation across responses is considered. By exploiting the correlation structure with a multivariate model, efficiency and power could be greatly increased (Marshall et al., 2006). The methodology proposed for parameter estimation in M-NLMMs is based on first-order (FO) approximations. Marshall et al, 2006 developed a Taylor series linear approximation to the marginal means and variance-covariance matrices of a M-NLMM, describing an EM-type algorithm for parameter estimation. Their EM algorithm extends the method described in Hirst et al. (1991) for parameter estimation in univariate (single response variable) NLMM using the FO approximation obtained by expanding the conditional mean about the average random effect. The latter proposal was later generalized by Young, Zerbe and Hay (1992) to include the Lindstrom and Bates (1990) FO approximation obtained by considering an expansion around the posterior mode of random effects. Also, Hall and Clutter (2004) used the EM algorithm based on linear approximations to estimate multivariate multilevel nonlinear mixed effects models.

The main objective of this article is to explore a classification technique for predicting class membership on the basis of a longitudinally measured multivariate response. The inference problem is formally described as a discriminant analysis based on a multivariate nonlinear mixed-effects model for longitudinal data. Additionally, missing data often occur in longitudinal studies because individuals miss some of their regular appointments or because some variables may not be measured at particular visits. Thus, we also consider the case where only a subset of the r responses may be observed at any occasion. Our approach provides the posterior probability of belonging to one of k classes based on having r different responses measured repeatedly over time. In the context of our motivating example, physicians would be able to make treatment or intervention decisions on the basis of these probabilities. Further, we consider a flexible and realistic data structure that allows dealing with joint modeling and classification for patients with very few or just one observation. This is especially relevant for the motivating example where 62 per cent of the patients had two or fewer measurements.

In our approach the classes or groups are predefined and the task is to understand the basis for the classification from a set of labeled units (training dataset). This information is then used to classify future units. In the case where classes or groups are unknown a priori and need to be estimated from the data, latent class models offer a fruitful approach. See additional developments along this direction in Lin et al., (2000), Muthén et al., (2002), and references therein.

The article is organized as follows. In Section 2 we describe the statistical methodology for the classification of multivariate longitudinal data using multivariate nonlinear mixed-effects models, briefly discussing EM-type algorithms for parameter estimation. We illustrate the proposed multivariate longitudinal method in Section 3 using data from Santiago, Chile on β-HCG and estradiol measured in women with normal and abnormal pregnancy outcomes. In Section 4 we evaluate the performance of the classification procedure with two estimation methods using simulated data. Finally, we summarize and discuss implications in Section 5.

2 Discriminant Analysis with Multivariate Longitudinal Data

2.1 Model specification

Suppose that, for the ith of m units, we observe data at n_i time points. At the jth time point, t_{ij} $(j = 1, \ldots, n_i)$, we have r continuous responses y_{ijk} $(k = 1, \ldots, r)$. Now, let $\boldsymbol{Y}_i = [\boldsymbol{y}_{i1}, \boldsymbol{y}_{i2}, \ldots, \boldsymbol{y}_{ir}]$ be the response matrix for unit i where $\boldsymbol{y}_{ik}$ is an $n_i \times 1$ vector response for variable k. Similarly, let $\boldsymbol{E}_i = [\boldsymbol{\epsilon}_{i1}, \boldsymbol{\epsilon}_{i2}, \ldots, \boldsymbol{\epsilon}_{ir}]$ be the matrix of error terms associated with $\boldsymbol{Y}_i$. Let $\boldsymbol{y}_i = \text{vec}(\boldsymbol{Y}_i)$ and $\boldsymbol{\epsilon}_i = \text{vec}(\boldsymbol{E}_i)$ denote a stacked $rn_i \times 1$ vector of all the response variables for unit i and error terms, respectively. Most stochastic models for serial measurements can be classified either as full multivariate or multi-stage random-effects models. In the full multivariate model, it is assumed that each vector of multiple responses $\boldsymbol{y}_i$, inside the ℓth of g groups or populations, is multivariate normal with mean $\boldsymbol{\mu}_{i\ell}(rn_i \times 1)$ and an arbitrary $rn_i \times rn_i$ dispersion matrix $\boldsymbol{R}_\ell$. The mean vector may depend upon the pattern of observations and also upon covariates.

When the design is balanced but observations are missing at random, traditional multivariate discriminant analysis based on the full multivariate model can be easily applied by using multivariate methods for missing observations (Dempster, Laird, and Rubin, 1977; Schafer, 1997). However, when units are measured at arbitrary or unique times, or when the dimension of $\boldsymbol{R}_\ell$ is large, this approach becomes unattractive, because a full multivariate model with unrestricted dispersion matrix requires a proliferation of variance parameters, many of which will be poorly estimated. In addition, the full multivariate model does not permit the definition and estimation of (random) unit-specific characteristics (Laird and Ware, 1982).

Two-stage random-effects models are based on explicit identification of unit-specific and population characteristics, and their form extends naturally to the unbalanced situation. The majority of work on methods for longitudinal data with multiple responses has focused on data that can be modeled by means of an expectation function that is assumed linear in its parameters (see Shah, Laird, and Schoenfeld, 1997; Schafer and Yucel, 2002; Fieuws and Verbeke, 2004; O'Brien and Fitzmaurice, 2005; and references therein). However, in many situations, we are concerned with data for which the assumption of normal errors is tenable but the proposed expectation function is nonlinear.

Let us consider in each group ℓ ($\ell = 1, \ldots, g$) a M-NLMM, that is, the model for the $rn_i \times 1$ multiple responses vector of the ith unit in group ℓ can be formulated as

$$\boldsymbol{y}_i = \boldsymbol{\mu}(\boldsymbol{\beta}_i, \mathbf{v}_i) + \boldsymbol{\epsilon}_i, \quad i = 1, \ldots, m, \tag{1}$$

where $\boldsymbol{\mu}$ is a nonlinear real-valued, differentiable function of a vector-valued mixed-effects parameter $\boldsymbol{\beta}_i$ and vector of covariates $\mathbf{v}_i$, and $\boldsymbol{\epsilon}_i$ is a vector containing the usual error components. The mixed-effects parameter $\boldsymbol{\beta}_i$ can be incorporated into the model as $\boldsymbol{\beta}_i = \boldsymbol{X}_i\boldsymbol{\beta}_\ell + \boldsymbol{Z}_i\boldsymbol{\beta}_i$, where $\boldsymbol{\beta}_i$ is a $q \times 1$ random-effects vector specific to the ith unit, and $\boldsymbol{Z}_i$ is the associated design matrix. The fixed-effects design matrix and parameter p-dimensional vector specific to the ℓ group are $\boldsymbol{X}_i$ and $\boldsymbol{\beta}_\ell$, respectively. We assume that $\boldsymbol{\beta}_i \sim \text{MVN}(0, \boldsymbol{B}_\ell)$ where $\boldsymbol{B}_\ell$ is a $q \times q$ positive-definite covariance matrix, $\boldsymbol{\epsilon}_i \sim \text{MVN}(0, \boldsymbol{R}_{i\ell})$. Here, $\boldsymbol{R}_{i\ell}$ is an $rn_i \times rn_i$ covariance matrix of the error terms that depends on i through its dimension rn_i but with a corresponding set of unknown parameters that does not. $\boldsymbol{B}_\ell$ is the covariance matrix of the random-effects in group ℓ and allows for covariance between the random-effects within a given response variable as well as covariance among the random-effects of different response variables. The $\boldsymbol{R}_{i\ell}$ covariance matrix in group ℓ has a specified structure to reflect the multivariate nature of the data. We assume that the observations at different time points are independent but that the multivariate responses at a particular time point are correlated with a $r \times r$ covariance matrix $\boldsymbol{\Sigma}_\ell$, which is the same for all time points. Consequently, $\text{Var}(\boldsymbol{\epsilon}_i) = \boldsymbol{R}_{i\ell} = \boldsymbol{\Sigma}_\ell \otimes \mathbf{I}_i$ where $\otimes$ denotes the Kronecker products. In this model, the marginal distribution of $\boldsymbol{y}_i$ can be difficult to find even in the case where the conditional distribution of $\boldsymbol{y}_i$ given $\boldsymbol{\beta}_i$ is normal and the marginal distribution of $\boldsymbol{\beta}_i$ is normal.

When the random effects are linear in the scale of the response vector the model can be formulated as

$$\boldsymbol{y}_i = \boldsymbol{\mu}(\boldsymbol{\beta}_\ell, \mathbf{v}_i) + \boldsymbol{Z}_i(\boldsymbol{\beta}_\ell)\boldsymbol{\beta}_i + \boldsymbol{\epsilon}_i. \tag{2}$$

where $\boldsymbol{Z}_i(\boldsymbol{\beta}_\ell)$ is a $rn_i \times q$ matrix of known functions of unknown parameters $\boldsymbol{\beta}_\ell$. In this case, the marginal distribution of the multiple response vector $\boldsymbol{y}_i$ is given by

$$\boldsymbol{y}_{i|\text{group}\,\ell} \sim \text{MVN}(\boldsymbol{\mu}_i(\boldsymbol{\beta}_\ell, \mathbf{v}_i), \boldsymbol{\Psi}_{i\ell}) \tag{3}$$

with $\boldsymbol{\Psi}_{i\ell} = \boldsymbol{Z}_i(\boldsymbol{\beta})\boldsymbol{B}_\ell\boldsymbol{Z}_i'(\boldsymbol{\beta}) + \boldsymbol{\Sigma}_\ell \otimes \mathbf{I}_i$ and the corresponding density is denoted by $f_{\ell i}(\boldsymbol{y}_i)$.

Given prior probabilities π_ℓ, $\ell = 1, \ldots, g$, for the g groups, and choosing a zero–one loss function (which minimizes the average error probability), the Bayes classification rule can be written:

$$\eta(\boldsymbol{y}_i) = \arg \max_{\ell=1,\ldots,g} p_{\ell i}(\boldsymbol{y}_i) \tag{4}$$

where $p_{\ell i}(\boldsymbol{y}_i)$ is the posterior probability of membership in group ℓ, i.e.,

$$p_{\ell i}(\boldsymbol{y}_i) = \frac{\pi_\ell f_{\ell i}(\boldsymbol{y}_i)}{\sum_s \pi_s f_{si}(\boldsymbol{y}_i)}. \tag{5}$$

Each unit is classified into the group for which the highest estimated posterior probability of membership is achieved. This is an optimal allocation rule based on the Neyman-Pearson lemma. Note that, although the M-NLMM (3) specifies the multiple response vector conditionally on a vector $\boldsymbol{\beta}_i$ of random effects, classification is based on the marginal distribution obtained from integrating over the random effects. In the case of model (1), where the random-effects $\boldsymbol{\beta}_i$ are part of the nonlinear component even in the case where these random parameters are normally distributed, the resulting marginal distribution of the multiple responses vector $\boldsymbol{y}_i$ is commonly unknown and difficult to find analytically. In that case, a linear approximation to the model residuals yields a marginal distribution of the unit-specific observations vector that is approximately normal, i.e.,

$$\boldsymbol{y}_{i|\text{group}\,\ell} \mathrel{\dot\sim} \text{MVN}(\boldsymbol{\mu}(\boldsymbol{\beta}_\ell, 0), \boldsymbol{\Psi}_{i\ell}), \tag{6}$$

with $\boldsymbol{\Psi}_{i\ell} = \tilde{\boldsymbol{Z}}_i(\boldsymbol{\beta}_\ell, 0)\boldsymbol{B}_\ell\tilde{\boldsymbol{Z}}_i'(\boldsymbol{\beta}_\ell, 0) + \boldsymbol{\Sigma}_\ell \otimes \mathbf{I}_i$, where $\tilde{\boldsymbol{Z}}_i(\boldsymbol{\beta}_\ell, 0)$ is the Jacobian matrix $\partial\boldsymbol{\mu}(\boldsymbol{\beta}_\ell, \boldsymbol{\beta}_i)/\partial\boldsymbol{\beta}_i$ evaluated at $\boldsymbol{\beta}_i = 0$.

In the remainder of this section the discussion will be based on a generic patient and thus we simplify the notation by dropping the subindex i. In classical discriminant analysis the Mahalanobis distance plays a central role in both the concep- tual framework and the allocation rules. The squared Mahalanobis distance between the multiple response vector $\boldsymbol{y}$ and the mean of the distribution of population ℓ, $\boldsymbol{\mu}_\ell$, with respect to $\boldsymbol{\Psi}_\ell$ is $D_\ell(\boldsymbol{y}) = (\boldsymbol{y} - \boldsymbol{\mu}_\ell)'\boldsymbol{\Psi}_\ell^{-1}(\boldsymbol{y} - \boldsymbol{\mu}_\ell)$. Having defined the Mahalanobis distance, the classification rule is to allocate $\boldsymbol{y}$ to population s if $\lambda_{\ell s}(\boldsymbol{y}) \leq 0$ for $\ell = 1, \ldots, g$ and $\ell \neq s$, where

$$\lambda_{\ell s}(\boldsymbol{y}) = D_s^*(\boldsymbol{y}) - D_\ell^*(\boldsymbol{y}) + 2\log\frac{\pi_\ell}{\pi_s}, \tag{7}$$

with $D_\ell^*(\boldsymbol{y}) = D_\ell(\boldsymbol{y}) + \log|\boldsymbol{\Psi}_\ell|$.

As elaborated in Marshall and Barón (2000), four discriminant models are possible according to the form of the variance of $\boldsymbol{y}$ in the population ℓ. Let θ denote the vector of all variance and covariance parameters found in $\boldsymbol{\Psi}$, that is, θ consists of the different elements in $\boldsymbol{B}$ and of all parameters in $\boldsymbol{\Sigma}$. The variance of $\boldsymbol{y}$ in the population ℓ, $\boldsymbol{\Psi}_\ell = \boldsymbol{\Psi}(\boldsymbol{\beta}_\ell, \theta_\ell) = \boldsymbol{Z}(\boldsymbol{\beta}_\ell)\boldsymbol{B}(\theta_\ell)\boldsymbol{Z}'(\boldsymbol{\beta}_\ell) + \boldsymbol{\Sigma}(\theta_\ell) \otimes \mathbf{I}_{n_i}$, is a function of the mean population-specific parameters $\boldsymbol{\beta}_\ell$ and the variance components θ_ℓ. For model (1), $\boldsymbol{Z}(\boldsymbol{\beta}_\ell) = \tilde{\boldsymbol{Z}}(\boldsymbol{\beta}_\ell, 0)$. The *homoscedastic model* is obtained when $\boldsymbol{Z}(\boldsymbol{\beta}_\ell) = \boldsymbol{Z}$ does not depend on the mean parameters $\boldsymbol{\beta}_\ell$ and the variance components are homogeneous, that is, $\theta_\ell = \theta$ for $\ell = 1, 2, \ldots, g$. In this situation $\boldsymbol{\Psi}_\ell = \boldsymbol{\Psi}$. In particular, when $g = 2$ the above classifier implies classification of a unit with multiple response vector $\boldsymbol{y}$ in the first group, if and only if

$$\{\boldsymbol{y} - (\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)/2\}'\boldsymbol{\Psi}^{-1}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) > \log(\pi_2/\pi_1)$$

which is the linear discriminant function.

The *mean-heteroscedastic model* consists of a model in which the design matrix $\boldsymbol{Z}(\boldsymbol{\beta}_\ell)$ depends on the mean parameters $\boldsymbol{\beta}_\ell$ but the variance components $\theta_\ell = \theta$ remain homogeneous among all populations $\ell = 1, 2, \ldots, g$. In the *variance-heteroscedastic model* the design matrix $\boldsymbol{Z}(\boldsymbol{\beta}_\ell)$ does not depend on the population parameters $\boldsymbol{\beta}_\ell$, but the variance components θ_ℓ are different across populations $\ell = 1, 2, \ldots, g$. This is the case where the between-unit variances $\boldsymbol{B}_\ell$ are different among the groups, or the within-unit variances $\boldsymbol{\Sigma}_\ell$ vary among the g populations, or both covariance matrices are different. The *fully-heteroscedastic model* consists of a model in which the design matrix $\boldsymbol{Z}(\boldsymbol{\beta}_\ell)$ depends on the population parameters $\boldsymbol{\beta}_\ell$, and the variance components θ_ℓ are different across the populations $\ell = 1, 2, \ldots, g$. We will perform our own comparison of these four models later in Section 3.

2.2 Classification with Missing Data

The above discussion assumes that all r responses are observed at each occasion, although the number and timing of observations may differ from unit to unit; that is, the data are complete in responses but unbalanced in number of observations per experimental unit. In clinical trials missing data are common, in which case only a subset of the r responses may be observed at any occasion. We will now show how the classification rule (7) can be easily adapted to cover the case of responses that are missing completely at random (Little and Rubin, 1987).

Let $\boldsymbol{y}_{new}$ denote the currently available partial responses vector for the new unit. Let $\boldsymbol{S}$ be the matrix of zeros and ones which 'selects' the elements of $\boldsymbol{y}$ which are actually observed; that is, the product $\boldsymbol{S}\boldsymbol{y}_{new} = \boldsymbol{y}^o_{new}$ gives the observed components of $\boldsymbol{y}_{new}$. If all components of $\boldsymbol{y}_{new}$ are observed, then $\boldsymbol{S}$ is the $rn^* \times rn^*$ identity matrix, where n^* is the number of measurements per response.

In the presence of missing data, the classification rule (7) implies allocating $\boldsymbol{y}_{new}$ to population s if $\lambda_{\ell s}(\boldsymbol{y}^o_{new}) \leq 0$ for $\ell = 1, \ldots, g$ and $\ell \neq s$, where

$$\lambda_{\ell s}(\boldsymbol{y}^o_{new}) = D^*_s(\boldsymbol{y}^o_{new}) - D^*_\ell(\boldsymbol{y}^o_{new}) + 2 \log \frac{\pi_\ell}{\pi_s}, \tag{8}$$

and

$$D^*_\ell(\boldsymbol{y}^o_{new}) = (\boldsymbol{y}^o_{new} - \boldsymbol{\mu}^o_\ell)' {\boldsymbol{\Psi}^o_\ell}^{-1} (\boldsymbol{y}^o_{new} - \boldsymbol{\mu}^o_\ell) + \log |\boldsymbol{\Psi}^o_\ell|,$$

with $\boldsymbol{\mu}^o_\ell = \boldsymbol{S}\boldsymbol{\mu}_\ell$, and $\boldsymbol{\Psi}^o_\ell = \boldsymbol{S}\boldsymbol{\Psi}_\ell\boldsymbol{S}'$.

2.3 Parameter Estimation

The algorithms proposed for computing the maximum likelihood estimates (MLE) of $(\boldsymbol{\beta}, \boldsymbol{\Sigma}, \boldsymbol{B})$ and empirical Bayes estimators (predictors) for the random effects $\boldsymbol{\beta}_i$ for M-NLMMs, rely on iteratively linearizing the conditional mean function and solving the resulting multivariate linear mixed-effects model (M-LMM). Marshall et al. (2006) extend the basic model to handle multivariate reponses and used an EM-type algorithm to estimate the model parameters, also considering the case when missing data are present. They also show how to implement a variant of their algorithm using SAS Proc NLMIXED with the first-order approximation method of Beal and Sheiner (1988), which instead of evaluating the current values of the random effects $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}$ at each iteration, evaluates instead at $\boldsymbol{\beta} = 0$. See details in Marshall et al. (2006). Also, SAS Proc NLMIXED performs exact MLE using numerical integration. In principle, this approach could be extended to the multivariate case using the trick implemented in Marshall et al. (2006). In this case, the marginal density of the observation vector can be obtained using Monte Carlo integration. Details for implementing this procedure are given in Section 4.

Minor modifications to the algorithm developed in Marshall et al. (2006) are necessary to estimate the multivariate models described at the end of Section 2.1. Specifically, if some or all components of $\boldsymbol{\beta}$ are hypothesized to be group-specific, the design matrix $\boldsymbol{X}$ in the linearized version of M-NLMM is appropriately changed. When the variance components $\boldsymbol{\Sigma}$ and $\boldsymbol{B}$ differ across groups, we estimate them separately. All of these models can be straightforwardly estimated using SAS Proc NLMIXED. Details are provided in the next Section.

3 Data Analysis

It is well known in obstetrics that β-HCG and estradiol are clinical variables that show dramatic changes in women during pregnancy (Yamashita et al., 1989). It has been also established that values of these variables are different in women who have normal pregnancies with terminal deliveries than in women who have spontaneous abortions or other types of adverse pregnancy outcomes. This association has made it possible to classify, with some degree of uncertainty, the outcome of pregnancy.

We consider here a follow-up study of 161 young women, representing different pregnancies over a period of 2 years in a private fertilization obstetrics clinic in Santiago, Chile. Estradiol and β-HCG concentrations on the 161 women were measured during the first trimester of pregnancy. One of the main objectives of the study was to evaluate these concentrations at early stages of pregnancy, with the purpose of identifying women with a high risk of loss. Consequently, pregnancy outcomes were divided into two groups: normal and abnormal. The women were classified as normal pregnancies if they had a normal delivery, or as abnormal pregnancies if they had any complication resulting in a non-terminal delivery and loss of the fetus. The resulting dataset consists of 124 patients diagnosed with normal pregnancy and 37 patients with abnormal pregnancy outcome.

The responses that we analyze are the vectors of time-varying estradiol and β-HCG measurements for the 161 women. The 161 women altogether contribute a total of 348 observations per response, where the number the samples per woman ranged from 1 through 6, with median 2. There was some missing data but there was no reason to believe that the missingness was non-ignorable. Throughout we assume that the missing responses are missing completely at random. Missingness rates for the responses for normal and abnormal groups are 3% and 0% for β-HCG and 27% and 58% for estradiol, respectively.

Figure 1 presents the patient-specific profiles of estradiol and β-HCG on the $\log_{10}$ scale for both groups. The two populations appear clearly distinct when considering the ensemble of profiles. However, for any one of these profiles the classification into one or the other subpopulation is far less certain, in particular when considering a series of partial responses. Our main inference goal in analyzing these data is to provide a classification rule for a new patient. The rule should allow sequential updating as data accrues for the new patient.

For the period of observation, roughly the first trimester of pregnancy, the mean gestational ages (days) in women with normal and abnormal pregnancies, were 34.5, and 32.6 days, respectively; thus, no significant difference in gestational age was found among the groups. We found that concentrations of estradiol and β-HCG were significantly lower in women with an abnormal pregnancy than in those with a normal pregnancy. In addition, we assumed prior probabilities of group membership to be proportional to the size of the groups in the training sample.

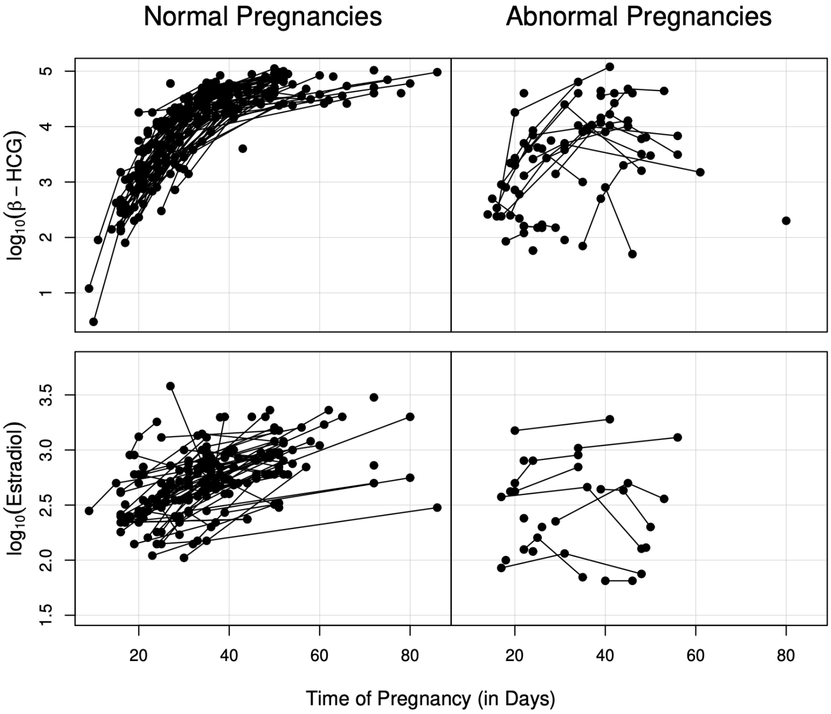
Marshall et al. (2006) proposed a nonlinear mixed and a linear mixed model to analyze the evolution of β-HCG and estradiol responses, respectively. The models were validated by fitting a series of more complex models and comparing them with respect to their $-2\log$ likelihoods. To test the differences between the two groups, they compared five alternatives and the best model among these five considered is the one having a different curve for each group and set of responses. We adopt the same model here. Letting $y_{i1\ell}$ and $y_{i2\ell}$ denote the β-HCG and estradiol responses, respectively, for patient i in group ℓ taken at time t, the models are specified as

$$y_{i1\ell}(t) = \frac{\beta_{\ell 1} + b_{i1\ell}}{1 + \exp\{(\beta_{\ell 2} - t)/\beta_{\ell 3}\}} + e_{i1\ell}(t) \tag{9}$$

$$y_{i2\ell}(t) = \beta_{\ell 4} + \beta_{\ell 5} t + b_{i2\ell} + e_{i2\ell}(t) \tag{10}$$

in which t is time expressed in days and $\ell = 1$ for the normal pregnancy and $\ell = 2$ for the abnormal pregnancy group. The vector $(\beta_{\ell 1}, \beta_{\ell 2}, \beta_{\ell 3}, \beta_{\ell 4}, \beta_{\ell 5})'$ of fixed effects describes the average evolution of the responses in group $\ell = 1, 2$ and the vector $(b_{i1\ell}, b_{i2\ell})$ of random effects describes

Fig. 1. Time profiles for normal and abnormal patients.



how the profile of the ith patient deviates from the average profiles in groups $\ell = 1, 2$. Both response trajectories are tied together through a joint distribution for the random effects

$$\boldsymbol{\beta}_{i\ell} = [b_{i1\ell}, b_{i2\ell}]' \sim \text{MVN}(0, \boldsymbol{B}_\ell),$$

where $\boldsymbol{B}_\ell$, the covariance matrix of the random effects, has elements given by $\left(B_{rs}^{(\ell)}\right)_{r=1,2;s=1,2}$. The error components are correlated and not associated with the random effects

$$\boldsymbol{\epsilon}_{i\ell} = [e_{i1\ell}, e_{i2\ell}]' \sim \text{MVN}(0, \boldsymbol{R}_{i\ell})$$

with $\boldsymbol{R}_{i\ell} = \boldsymbol{\Sigma}_\ell \otimes \mathbf{I}_{n_i}$, and $\boldsymbol{\Sigma}_\ell$ has elements $\left(\Sigma_{rs}^{(\ell)}\right)_{r=1,2;s=1,2}$. The latter implies that conditional on the random effects, the response trajectories are dependent. Special cases can be obtained by making additional assumptions about the covariance matrix $\boldsymbol{B}_\ell$ and $\boldsymbol{\Sigma}_\ell$. For example, if $B_{12}^{(\ell)}$ and $\Sigma_{12}^{(\ell)}$ are all equal to zero, the responses in each group are assumed to be completely independent at any point in time. In that case, parameters of the models can be obtained using likelihood based inference with, for example, the `NLME` software of Pinheiro and Bates (2000).

The marginal distribution of the response vector $\boldsymbol{y}_i$ in the ℓth group is

$$\boldsymbol{y}_i \sim \text{MVN}\left(\boldsymbol{\mu}_{i\ell}(t_i), \tilde{\boldsymbol{Z}}_{i\ell}\boldsymbol{B}_\ell\tilde{\boldsymbol{Z}}_{i\ell}' + \boldsymbol{\Sigma}_\ell \otimes \mathbf{I}_{n_i}\right) \tag{11}$$

where the mean vector $\boldsymbol{\mu}_{i\ell}(t_i)$ of dimension $2n_i \times 1$ has elements

$$\mu_{i\ell}(t_{ij}) = \begin{pmatrix} \beta_{\ell 1}/(1 + \exp\{(\beta_{\ell 2} - t_{ij})/\beta_{\ell 3}\}) \\ \beta_{\ell 4} + \beta_{\ell 5} t \end{pmatrix} \tag{12}$$

and represents the population curve at time t_{ij}, and $\tilde{\boldsymbol{Z}}_{i\ell}$ is a $2n_i \times 2$ working design matrix with rows made up of $\tilde{\boldsymbol{z}}_{ij\ell}$ where

$$\tilde{\boldsymbol{z}}_{ij\ell} = \begin{pmatrix} v_{ij\ell} & 0 \\ 0 & 1 \end{pmatrix}$$

with $v_{ij\ell} = 1/(1 + \exp\{(\beta_{\ell 2} - t_{ij})/\beta_{\ell 3}\})$ which depends on the values of the unknown population parameters $\beta_{\ell 2}$ and $\beta_{\ell 3}$.

Expressions (11) and (12) constitute a fully-heteroscedastic model. If desired, a mean-heteroscedastic model can be obtained from expression (11) by introducing the restrictions $\boldsymbol{B}_1 = \boldsymbol{B}_2$ and $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$, and the variance-heteroscedastic model from expression (12) by introducing the restrictions $\beta_{12} = \beta_{22}$ and $\beta_{13} = \beta_{23}$. The homoscedastic model is obtained by introducing both sets of restrictions.

We used the `NLME` software of Pinheiro and Bates (2000) to fit the models for each response separately to obtain initial estimates of model parameters. In this example, the better approximation is to expand about 0 since this yields the correct marginal mean and variance-covariance. If one uses the EM algorithm expanding about the posterior mode, there can be some bias as the result of using the derivative matrix evaluated at the posterior mode rather than at 0 (see Vonesh and Chinchilli, 1997). Thus, we used SAS Proc NLMIXED (method=firo) to estimate the final models.

The results of fitting the four models are shown in Table 1. Likelihood ratio tests comparing the more restricted models to the fully-heteroscedastic model show that the differences in β-HCG and estradiol between the groups occur not only at the mean level but also at the within-group and patient-specific levels. From Table 1 we conclude that the fully-heteroscedastic model is the best for discriminating between normal and non-terminal deliveries.

Parameter estimates for the four models are shown in Table 1. Observed differences across the groups in the linear and logistic curve parameters reflect the changes in $-2\log L$ between, principally, the homoscedastic and mean-heteroscedastic models. Observed differences in the group-specific estimates of the variance components in the variance- and fully-heteroscedastic models suggest that there is significantly more between-patient variability in the abnormal group than in the normal group.

Table 2 (part [A]) presents the classification results using the original sample. Seven of the 124 women having normal pregnancy were classified as abnormal whereas 16 of the 37 women having abnormal pregnancy were classified as normal. The observed misclassification rate is 14.3% (23/161). Other quantities of interest can be readily evaluated. Among these, "sensitivity" and "specificity" are popular ways to summarize the classification results. Letting $A = \{$ the patient is classified as abnormal $\}$, and $I = \{$ the patient actually belongs to the abnormal group $\}$, then sensitivity and specificity are defined respectively as $\mathbb{P}[A|I]$ and $\mathbb{P}[\bar{A}|\bar{I}]$. From our results we found 56.8% sensitivity and 94.4% specificity.

It is well known that the error rate obtained by applying the classifier to the same data from which it has been formed tends to be biased downward as an estimate of the true error rate. For this reason we used the leave-one-out cross-validation to obtain more accurate estimates of the misclassification rate. Table 2 (part [B]) presents the corresponding results, with 24 women and an estimated misclassification rate of 14.9%.

The Receiver Operating Characteristic (ROC) curves and the area under the ROC curve (AUC) for one multivariate and two univariate models are presented in Figure 2. Specifically, we present three curves showing the changes in sensitivity and specificity using only the estradiol variable, only the β-HCG variable, and using both the estradiol and β-HCG variables. Using the bivariate responses improves the sensitivity and specificity for predicting an abnormal pregnancy outcome in this population of women, a result that is expected via the Neyman-Pearson lemma for comparison of the classification performance of a multivariate likelihood ratio based allocation rule vs. the best of the univariate likelihood ratio-based allocation rules (Pepe, 2003).

In the above analysis we have used all of the available information. However, it is interesting to assess the predictive power of our model as a function of the number of observations used or their timing. Thus we generated from the corresponding fitted distributions, one future patient for each group and evaluated the classifier defined through equations (4) and (5) for up to six possible observations. Time points were randomly chosen from the empirical distribution of observed times within each group. Figure 1 of Web Appendix A shows the evolution of posterior probabilities for classifying one normal and one abnormal pregnancy outcome patient in the future. For the normal patient, we observe a steady growth of the probabilities. For the abnormal patient, this probability decreases to values that leave no question about the classification. It is interesting to note that when using only the β-HCG profiles, the classification probabilities for abnormal patients require more observations than the normal ones to achieve correct classification (see De la Cruz-Mesía and Quintana, 2007; De la Cruz-Mesía et al., 2007).

4 Simulation Study

In this section we report the results of a simulation study motivated by a problem related to the pregnant women example. We investigate the impact of two estimation methods on the resulting

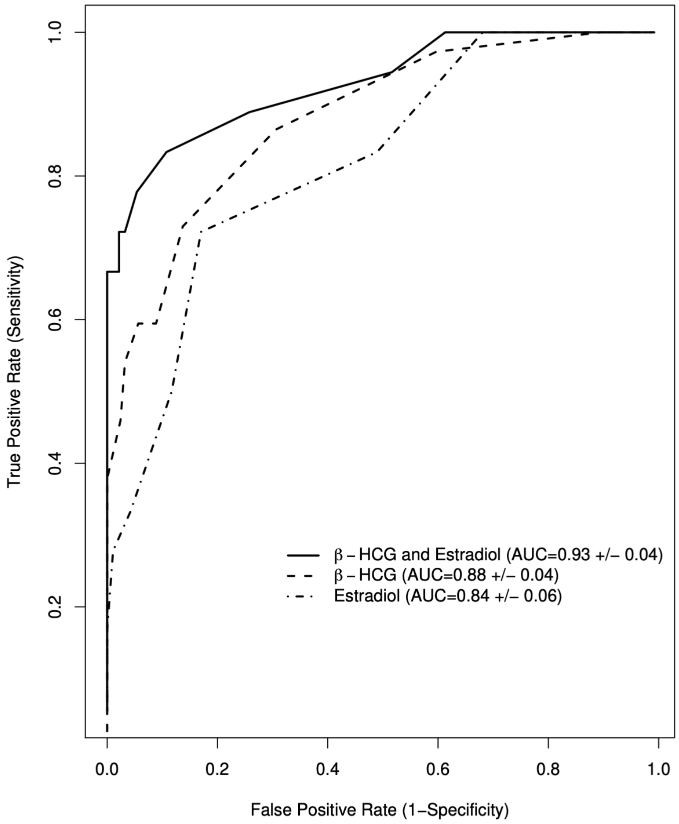
Table 1. Parameter estimates and standard errors (in parentheses) for the discriminant models.

	Estimates of Model			
	Homoscedastic	Heteroscedastic		
Parameter		Mean	Variance	Fully
β_{11}	$4.715_{(0.064)}$	$4.785_{(0.068)}$	$4.725_{(0.046)}$	$4.739_{(0.047)}$
β_{12}	$14.923_{(0.415)}$	$15.710_{(0.405)}$	$15.348_{(0.342)}$	$15.616_{(0.340)}$
β_{13}	$7.578_{(0.538)}$	$7.731_{(0.549)}$	$7.424_{(0.431)}$	$7.385_{(0.432)}$
β_{14}	$2.276_{(0.047)}$	$2.261_{(0.046)}$	$2.274_{(0.045)}$	$2.271_{(0.045)}$
β_{15}	$0.013_{(0.001)}$	$0.013_{(0.001)}$	$0.013_{(0.001)}$	$0.013_{(0.001)}$
β_{21}	$3.930_{(0.094)}$	$3.649_{(0.107)}$	$3.938_{(0.163)}$	$3.673_{(0.182)}$
β_{22}	–	$12.139_{(1.344)}$	–	$12.406_{(1.792)}$
β_{23}	–	$6.233_{(1.253)}$	–	$6.535_{(1.917)}$
β_{24}	$2.409_{(0.117)}$	$2.482_{(0.112)}$	$2.389_{(0.126)}$	$2.481_{(0.124)}$
β_{25}	$0.001_{(0.003)}$	$-0.001_{(0.003)}$	$0.002_{(0.003)}$	$-0.001_{(0.003)}$
$B_{11}^{(1)}$	$0.151_{(0.032)}$	$0.164_{(0.031)}$	$0.030_{(0.013)}$	$0.031_{(0.013)}$
$B_{22}^{(1)}$	$0.050_{(0.010)}$	$0.052_{(0.010)}$	$0.040_{(0.008)}$	$0.040_{(0.008)}$
$B_{12}^{(1)}$	$0.036_{(0.016)}$	$0.041_{(0.017)}$	$0.011_{(0.009)}$	$0.012_{(0.009)}$
$B_{11}^{(2)}$	–	–	$0.721_{(0.237)}$	$0.651_{(0.199)}$
$B_{22}^{(2)}$	–	–	$0.086_{(0.035)}$	$0.086_{(0.035)}$
$B_{12}^{(2)}$	–	–	$0.110_{(0.080)}$	$0.103_{(0.075)}$
$\Sigma_{11}^{(1)}$	$0.136_{(0.015)}$	$0.119_{(0.013)}$	$0.094_{(0.011)}$	$0.093_{(0.011)}$
$\Sigma_{22}^{(1)}$	$0.028_{(0.004)}$	$0.028_{(0.004)}$	$0.028_{(0.004)}$	$0.028_{(0.004)}$
$\Sigma_{12}^{(1)}$	$0.014_{(0.007)}$	$0.010_{(0.006)}$	$0.004_{(0.006)}$	$0.003_{(0.006)}$
$\Sigma_{11}^{(2)}$	–	–	$0.235_{(0.053)}$	$0.195_{(0.044)}$
$\Sigma_{22}^{(2)}$	–	–	$0.029_{(0.010)}$	$0.027_{(0.009)}$
$\Sigma_{12}^{(2)}$	–	–	$0.038_{(0.018)}$	$0.032_{(0.015)}$
	Summary of model fitting			
d.f.	14	16	20	22
-2 log L	425.2	403.0	319.6	310.1
AIC	453.2	435.0	359.6	354.1
χ^2	115.1	92.9	9.5	–
p-value	< 0.01	< 0.01	<0.01	-

Table 2. Classification results within-sample **(A)** and using Cross-validation **(B)**

	Classification				
	(A)		**(B)**		
Groups	Normal	Abnormal	Normal	Abnormal	
Normal	117	7	117	7	124
Abnormal	16	21	17	20	37
	133	28	134	27	161

Fig. 2. ROC curves with areas and standard errors for three discriminant models.



parameter estimates and their influence on the performance of the classification rule. The effects of number of patients and number of measurements per patient are examined. In our example we have only one random effect in model (9), entering the model in a strictly linear fashion. The same is true for the random effect in model (10). Thus, the mean curves coincide with the population-averaged curves (see eq. 11). It is natural, however, to consider the case where the random effects enter the models in a nonlinear fashion, so that the impact of the above approximation is far less clear on estimation and particularly on our classification rule. We describe next a specific scheme that will shed some light on these (and other) issues.

In the example we have 124 patients in the normal pregnancy group with the number of measurements per patient ranging from 1 to 4 (median 2) and in the abnormal group we have 37 patients with the number of measurements per patient ranging from 1 to 6 (median 2). We judge the number of observations in each group to be sufficiently large.

Our simulations use two groups, with 124 and 37 patients, respectively. We consider three cases for the number of measurements per patient within each group. The first corresponds to the situation described in the example, the second has five ($n_i = 5$) measurements, and the last ten ($n_i = 10$) measurements, respectively per patient. A logistic model similar to (9) was used to generate one response, but with random effects entering in a nonlinear fashion. The other response was generated using model (10). To model the simulated data we use in all cases

$$\begin{aligned} y_{i1\ell}(t) &= \frac{\beta_{\ell 1}}{1+\exp\{(\beta_{\ell 2}+b_{i2\ell}-t)/\beta_{\ell 3}\}} + e_{i1\ell}(t) \\ y_{i2\ell}(t) &= \beta_{\ell 4} + \beta_{\ell 5} t + b_{i4\ell} + e_{i2\ell}(t). \end{aligned}$$

Assuming a set of values for the model parameters and the group membership probabilities, a training data set is simulated. On this data set, parameter estimation for all models is carried out using the first order and Gaussian quadrature methods. Next, we simulate a validation data set on which we run the following two classification procedures: *(i)* using the MLEs obtained by Gaussian quadrature we implement the Bayes classifier defined through equations (4) and (5). This procedure involves evaluating the density of the observation vector by Monte Carlo integration with a large number of draws (we chose 10 000) from the random effects distribution. This is done only once per subject. *(ii)* using the estimates obtained with the first order estimation method, the Bayes classifier is implemented using density (6). The entire process is replicated 1 000 times.

Tables 3 and 4 present the estimation results using the FO approximation and Gaussian quadrature. The column labeled “Estimate” denotes the average values the parameter estimates over 1 000 replications; “SE” is the standard error of estimates. Our simulation results indicate that the two approximations produce similar estimates of the fixed effects $\boldsymbol{\beta}$ but yield some differences for the $\boldsymbol{B}$ parameters in the abnormal group with unbalanced number of measurements per patient. The FO approximation grossly overestimates B_{11}, the variance of b_{i2}, and underestimates B_{12}. This is probably because the precision in the estimation of $\boldsymbol{\Sigma}$ is determined by the total number of observations, but the precision of the estimates of $\boldsymbol{B}$ is determined by the number of subjects. When the number of patients is 124, the estimates from the FO and Gaussian approximations were almost unbiased and very similar. The FO approximation, particularly when the number of patients is 37, showed considerably more variation than in the Gaussian case.

We turn now to the assessment of the classification rule. For this, we calculated the misclassification error rate, sensitivity and specificity. The first two columns of Figures 2 and 3 of Web Appendix A shows the results obtained using both procedures (each box in the figures shows the

Table 3. Simulation results of the parameters estimates using first-order (FO) and Gaussian quadrature methods considering 124 patients and a different number of measurements per patient (n_i).

True value	n_i	First Order			Gaussian quadrature		
		Estimate	SE	Bias	Estimate	SE	Bias
$\beta_{11} = 4.8$	$n_i = 4^*$	4.804	0.034	0.004	4.801	0.035	0.001
	$n_i = 5$	4.806	0.020	0.006	4.801	0.021	0.001
	$n_i = 10$	4.805	0.014	0.005	4.800	0.014	0.000
$\beta_{12} = 15.1$	$n_i = 4^*$	15.130	0.489	0.030	15.038	0.456	-0.062
	$n_i = 5$	15.159	0.353	0.059	15.089	0.391	-0.011
	$n_i = 10$	15.144	0.308	0.044	15.065	0.447	-0.035
$\beta_{13} = 7.6$	$n_i = 4^*$	7.837	0.396	0.237	7.643	0.363	0.043
	$n_i = 5$	7.824	0.226	0.224	7.614	0.224	0.014
	$n_i = 10$	7.833	0.156	0.233	7.628	0.163	0.027
$\beta_{14} = 2.3$	$n_i = 4^*$	2.302	0.037	0.002	2.324	0.111	0.024
	$n_i = 5$	2.299	0.025	-0.001	2.313	0.065	0.013
	$n_i = 10$	2.299	0.023	-0.001	2.316	0.050	0.016
$\beta_{15} = 0.01$	$n_i = 4^*$	0.010	0.001	0.000	0.010	0.002	0.000
	$n_i = 5$	0.010	5×10^{-4}	0.000	0.010	0.001	0.000
	$n_i = 10$	0.010	4×10^{-4}	0.000	0.010	7×10^{-4}	0.000
$\Sigma_{11} = 0.05$	$n_i = 4^*$	0.050	0.005	0.000	0.050	0.005	0.000
	$n_i = 5$	0.050	0.003	0.000	0.050	0.002	0.000
	$n_i = 10$	0.050	0.002	0.000	0.051	0.002	0.001
$\Sigma_{22} = 0.03$	$n_i = 4^*$	0.030	0.003	0.000	0.030	0.004	0.000
	$n_i = 5$	0.030	0.002	0.000	0.030	0.002	0.000
	$n_i = 10$	0.030	0.001	0.000	0.030	0.001	0.000
$\Sigma_{12} = -9 \times 10^{-4}$	$n_i = 4^*$	-7×10^{-4}	0.003	2×10^{-4}	-4×10^{-4}	0.002	5×10^{-4}
	$n_i = 5$	-5×10^{-4}	0.001	4×10^{-4}	-3×10^{-4}	0.001	6×10^{-4}
	$n_i = 10$	-6×10^{-4}	0.001	3×10^{-4}	-3×10^{-4}	0.001	6×10^{-4}
$B_{11} = 8.5$	$n_i = 4^*$	8.537	1.853	-0.037	8.292	0.667	-0.208
	$n_i = 5$	8.500	1.482	0.000	8.127	0.627	-0.373
	$n_i = 10$	8.398	0.888	-0.102	7.965	0.548	-0.535
$B_{22} = 0.04$	$n_i = 4^*$	0.041	0.008	0.001	0.040	0.007	0.000
	$n_i = 5$	0.040	0.006	0.000	0.040	0.006	0.000
	$n_i = 10$	0.040	0.005	0.000	0.038	0.006	-0.002
$B_{12} = -0.21$	$n_i = 4^*$	-0.213	0.103	-0.003	-0.204	0.089	0.006
	$n_i = 5$	-0.213	0.067	-0.003	-0.204	0.066	0.006
	$n_i = 10$	-0.208	0.057	0.002	-0.198	0.072	0.012

The notation 4^* refers to values of n_i, ranging from 1 to 4, exactly as in the pregnancy dataset.

median, quartiles, and extreme values within a category). All of these quantities are expressed as percentages. The results show that the number of measurements per patient does affect classification. This is not surprising because increasing the number of observations leads to a more accurate discrimination function and thus to a better classification procedure. For this reason, we only show the results when the number of measurements per patient is unbalanced and when $n_i = 5$. In view of the information provided by these figures, we did not find convincing evidence that the estimation procedures have any significant effect on classification.

It is well known that as $\mathbf{B} \to 0$, the FO method will still yield a reasonably unbiased estimate of both the population parameter and the population mean response (see, e.g., 78 Biometrics, March 2009 Vonesh and Chinchilli, 1997, p. 352–357; Demidenko, 2004, p. 455–462). This is what may be happening in the results of Tables 3 and 4, namely that the variance component $\mathbf{B}$ may be small enough to make the two approximations appear to behave in a very similar fashion. From Table 3, the coefficient of variation (CV) for the nonlinear random effect is of order 0.193 for the normal group ($\beta_{22} = 11.6$,$B_{11} = 1.03$). This is moderately small but not too small. Likewise, in Table 4, the CV associated with the nonlinear random effect is 0.0875 and this is considerably smaller. In essence, when B11 is “small,” all the random effects will be clustered near 0 and one can effectively approximate the likelihood by expanding around bi2 = 0. To show other situations, we carried out more simulations considering larger values for the CV associated with the random effects. In doing so, however, we note that the previous simulation results suggest that, in general, the exact Gaussian quadrature method yields reasonable estimation results. Therefore, in what follows we restrict ourselves to studying the behavior of the FO approximation method.

Tables 1 and 2 of Web Appendix A show the results ob- tained when simulating samples with various CV values. We can clearly see that when the CV increases, the linearization approximation involved in the FO method gets progressively worse, especially when we consider 37 patients. In contrast with our previous simulation results, we note that increasing the CV has also a negative effect on the classification procedures (see Figures 2 and 3 of Web Appendix A) basically because of the large variability of the nonlinear random effects.

In summary, our results suggest that the likelihood approximation using the FO method performs well when the number of the intraunit measurements is not small and the variability of the random effects is not large. But when some of the units have either sparse data or a large variability of the associated random effects, considerable errors are introduced when using the FO method. This may explain why the results of replicating the observed sample sizes from the pregnancy data (denoted as $ni = 6^*$ in Table 4 and Table 2 of Web Appendix A) differ so much from those obtained when fixing $n_i = 5$.

5 Comments and Conclusions

The main purpose of this article was to propose an analytic approach to the development of discriminant functions when multiple responses are measured over time. Parameter estimates from nonlinear random-effects models for multiple responses are embedded into linear and quadratic discriminant models that are functions of time. We also consider the case when only a subset of the responses may be observed at any occasion.

Using simulation we compared the classification rule based on the squared Mahalanobis distance function, assuming the population mean vectors and variance–covariance matrices for g groups of units, with that obtained under Taylor series linear approximations to the marginal means and variance–covariance matrices. We found no significant differences. Our simulation study also suggests that the ML estimates based on the FO approximation are similar to “exact” ML estimates obtained using Gaussian quadrature when the interunit variability is small. From the simulation results we can see that the FO method performs poorly unless the ni are sizeable and the time points are clustered in the same region of the design space. Otherwise, large variance components may lead to biased or inconsistent estimates of the parameters, and more generally, to a bad performance of the discrimination based on this approach.

It is also worth noting that the proposed procedure involves two Taylor approximations, one at the parameter estimation stage and the other at the classification stage. We explore the effect of each of these approximations on the results separately. When parameter estimation was carried out using exact MLE with the Gaussian quadrature, but the classification was done using the Taylor approximation, similar results were obtained only for lower values of CV (see Figures 4 and 5 of Web Appendix A).

In summary, we find the linearization approach to be most useful when we have a CV of 20% or less, because there is little difference with respect to the quadrature Gaussian approach. On the contrary, the method appears to break down for large random-effects variances. The principal advantage of these discriminant models for unbalanced repeated measures is their ability to use all of the available information for classifying units over time, regardless of the number or timing of the observations. An area of concern and interest in longitudinal data analysis, but not pursued in this research, is the subject of informative censoring. This occurs when the response trajectory, or outcome, and the censoring mechanism for follow-up are not independent of each other. Although the data from our motivating example did not appear to exhibit informative censoring, the general setting from which the study arose could have led to shorter follow-up times on average for the women with ab- normal pregnancy outcomes. In that case, the random effects would be correlated with follow-up time. Further research is called for to investigate the effects of informative censoring on parameter estimates and classification using our proposed models.

A second advantage of these longitudinal discriminant models is that the influence of both between-group variability and components of within-group variability on discrimination can be readily quantified. In the case of a single response, the discriminant model reduces to that proposed by Marshall and Bar ́ on (2000). The multivariate model takes into account the correlation between responses at the same visit, and so it can discriminate much better than its univariate competitors. The results obtained from models incorporating additional biological and clinical information should help both physicians and patients to make more informed treatment decisions on the basis of objective staging data. Indeed, using the tools presented here, a reliable and inexpensive diagnostic test for early differentiation between normal and adverse outcome might re- duce the psychological tension and anxiety present in many patients. For patients judged by this test to be at high risk of an unfavorable outcome, a more careful follow-up might lead to better patient management and interventions that reduce risk.

Finally, an extension of the discriminant procedure that would be of particular interest in many applications is the modeling of group membership probabilities as a function of covariates. In the context of our example, this could be accom- plished using a logistic form for the population proportion of abnormal pregnancies, $\pi = \exp(\boldsymbol{\alpha}'\mathbf{x})/(1 + \exp(\boldsymbol{\alpha}'\mathbf{x}))$, where $\mathbf{x}$ is a vector of covariates for each women and $\boldsymbol{\alpha}$ a vector of regression parameters. Identifying associations between women’s covariates, such as age, number of previous normal and abnormal pregnancies, smoking status, and normal pregnancy tendencies, can be useful for targeting specific units in future analysis. Also the model for y i could depend on these covariates. In our example a number of women had missing covariate values.

Table 4. Simulation results of the parameter estimates using first-order (FO) and Gaussian quadrature methods considering 37 patients and a different number of measurements per patient (n_i).

		First Order			Gaussian quadrature		
True value	n_i	Estimate	SE	Bias	Estimate	SE	Bias
$\beta_{21} = 3.9$	$n_i = 6^*$	3.965	0.308	0.065	3.953	0.126	0.053
	$n_i = 5$	3.928	0.133	0.028	3.950	0.090	0.050
	$n_i = 10$	3.918	0.085	0.018	3.945	0.053	0.047
$\beta_{22} = 11.6$	$n_i = 6^*$	11.945	2.800	0.345	11.875	0.881	0.275
	$n_i = 5$	11.543	1.470	-0.057	11.870	0.705	0.270
	$n_i = 10$	11.586	1.049	-0.014	11.952	0.300	0.352
$\beta_{23} = 8.5$	$n_i = 6^*$	9.177	4.240	0.678	8.958	0.784	0.458
	$n_i = 5$	8.854	1.442	0.354	8.947	0.512	0.447
	$n_i = 10$	8.746	1.027	0.246	8.987	0.249	0.487
$\beta_{24} = 2.3$	$n_i = 6^*$	2.291	0.100	-0.009	2.137	0.095	-0.163
	$n_i = 5$	2.298	0.076	-0.002	2.128	0.087	-0.172
	$n_i = 10$	2.298	0.064	-0.002	2.108	0.069	-0.192
$\beta_{25} = 0.002$	$n_i = 6^*$	0.002	0.003	0.000	0.005	0.002	0.003
	$n_i = 5$	0.002	0.001	0.000	0.005	0.001	0.003
	$n_i = 10$	0.002	0.001	0.000	0.005	0.001	0.003
$\Sigma_{11} = 0.61$	$n_i = 6^*$	0.575	0.094	-0.035	0.595	0.089	-0.015
	$n_i = 5$	0.600	0.064	-0.010	0.603	0.063	-0.007
	$n_i = 10$	0.608	0.045	-0.002	0.611	0.045	0.001
$\Sigma_{22} = 0.05$	$n_i = 6^*$	0.048	0.009	-0.002	0.047	0.009	-0.003
	$n_i = 5$	0.050	0.006	0.000	0.049	0.006	-0.001
	$n_i = 10$	0.050	0.004	0.000	0.050	0.004	0.000
$\Sigma_{12} = 0.12$	$n_i = 6^*$	0.113	0.025	-0.007	0.115	0.025	-0.005
	$n_i = 5$	0.118	0.017	-0.002	0.118	0.017	-0.002
	$n_i = 10$	0.119	0.011	-0.001	0.119	0.011	-0.001
$B_{11} = 1.03$	$n_i = 6^*$	2.948	9.137	1.918	1.002	0.231	0.028
	$n_i = 5$	1.095	1.158	0.065	1.007	0.293	-0.023
	$n_i = 10$	0.996	0.798	-0.034	0.991	0.127	-0.039
$B_{22} = 0.1$	$n_i = 6^*$	0.096	0.023	-0.004	0.103	0.032	0.003
	$n_i = 5$	0.095	0.022	-0.005	0.100	0.028	0.000
	$n_i = 10$	0.092	0.021	-0.008	0.098	0.025	-0.002
$B_{12} = -0.23$	$n_i = 6^*$	-0.062	0.234	0.168	-0.208	0.078	0.022
	$n_i = 5$	-0.221	0.370	0.009	-0.226	0.070	0.004
	$n_i = 10$	-0.220	0.262	0.010	-0.256	0.054	-0.026

$6^* \equiv 1 \leq n_i \leq 6$.

Bibliography

[1] Beal, S. L. and Sheiner, L. B. (1988). Heteroskedastic nonlinear regression. Technometrics 30, 327-338.

[2] Brant, L. J., Sheng, S. L., Morrell, C. H., Verbeke, G. N., Lesaffre, E., and Carter, H. B. (2003). Screening for prostate cancer by using random-effects models. Journal of the Royal Statistical Society, Series A 166, 51–62.

[3] Brown, P. J., Kenward, M. G., and Bassett, E. E. (2000). Bayesian discrimination with longitudinal data. Biostatistics 2, 417– 432.

[4] De la Cruz-Mesía, R. and Quintana, F. A. (2007). A model-based approach to Bayesian classification with applications to predicting pregnancy outcomes from longitudinal β-hCG profiles. Biostatistics 8, 228–238.

[5] De la Cruz-Mesía, R.,Quintana, F.A.,and Müller, P.(2007). Semipara- metric Bayesian classification with longitudinal markers. Journal of the Royal Statistical Society, Series C, (Applied Statistics) 56, 119–137.

[6] Demidenko, E. (2004). Mixed Models: Theory and Applications. New York: Wiley.

[7] Dempster, A. E., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood with incomplete data via the E-M algorithm. Journal of the Royal Statistical Society, Series B 39, 1–38.

[8] Fieuws, S. and Verbeke, G. (2004). Joint modelling of multivariate lon- gitudinal profiles: Pitfalls of the random-effects approach. Statis- tics in Medicine 23, 3093–3104.

[9] Hall, D. B. and Clutter, M. (2004). Multivariate multilevel nonlinear mixed effects models for timber yield predictions. Biometrics 60, 16–24.

[10] Hirst, K., Zerbe, G. O., Boyle, D. W., and Wilkening, R. B. (1991). On nonlinear random effects models for repeated measurements. Communications in Statistics B, Simulation and Computation 20, 463–478.

[11] Laird, N. M. and Ware, J. H. (1982). Random effects models for longi- tudinal data. Biometrics 38, 963–974.

[12] Lin, H. Q., McCulloch, C. E., Turnbull, B. W., Slate, E. H., and Clark, L. C. (2000). A latent class mixed model for analysing biomarker trajectories with irregularly scheduled observations. Statistics in Medicine 19, 1303–1318.

[13] Lindstrom, M. J. and Bates, D. M. (1990). Nonlinear mixed effects models for repeated measures data. Biometrics 46, 673–687.

[14] Little, R. J. A. and Rubin, D. B. (1987). Statistical Analysis with Missing Data. New York: Wiley.

[15] Marshall, G. and Barón, A. E. (2000). Linear discriminant models for unbalanced longitudinal data. Statistics in Medicine 19, 1969–1981.

[16] Marshall, G., De la Cruz-Mesía, R., Barón, A. E., Rutledge, J. H., and Zerbe, G. O. (2006). Nonlinear random effects model for multivariate responses with missing data. Statistics in Medicine 25, 2817–2830.

[17] Muthén, B., Brown, C. H., Masyn, K., Jo, B., Khoo, S. T., Yang, C. C., Wang, C. P., Kellam, S. G., Carlin, J. B., and Liao, J. (2002). General growth mixture modeling for randomized preventive interventions. Biostatistics 3, 459–475.

[18] O'Brien, L. M. and Fitzmaurice, G. M. (2005). Regression models for the analysis of longitudinal Gaussian data from multiple sources. Statistics in Medicine 24, 1725–1744.

[19] Pepe, M. S. (2003). The Statistical Evaluation of Medical Tests for Classification and Prediction. NewYork: Oxford University Press.

[20] Pinheiro, J. C. and Bates, D. M. (2000). Mixed-Effects Models in S and S-PLUS. New York: Springer.

[21] Schafer, J. L. (1997). Analysis of Incomplete Multivariate Data. London: Chapman and Hall.

[22] Schafer, J. L. and Yucel, R. M. (2002). Computational strategies for multivariate linear mixed-effects models with missing values. Journal of Computational and Graphical Statistics 11, 437– 457.

[23] Shah, A., Laird, N., and Schoenfeld, D. (1997). A random-effects model for multiple characteristics with possibly missing data. Journal of the American Statistical Association 92, 775–779.

[24] Shapiro, B. S., Escobar, M., Makuch, R., Lavy, G., and DeCherney, A. H. (1992). A model-based prediction for transvaginal ultrasonagraphic identification of early intrauterine pregnancy. American Journal of Obstetrics and Gynecology 166, 1495–1500.

[25] Tomasko, L., Helms, R. W., and Snapinn, S. M. (1999). A discriminant analysis extension to mixed models. Statistics in Medicine 18, 1249-1260.

[26] Verbeke, G. and Lesaffre, E. (1996). A linear mixed-effects model with heterogeneity in the random effects population. Journal of the American Statistical Association 91, 217-221.

[27] Verbeke, G. and Molenberghs, G. (2000). Linear Mixed Models for Longitudinal Data. New York: Springer-Verlag.

[28] Vonesh, E. F. and Chinchilli, V. M. (1997). Linear and Nonlinear Models for the Analysis of Repeated Measurements. New York: Marcel Dekker, Inc.

[29] Wernecke, K.-D., Kalb, G., Schink, B., and Wegner, B. (2004). A mixed model approach to discriminant analysis with longitudinal data. Biometrical Journal 46, 246–254.

[30] Yamashita, T., Okamoto, S., Thomas, A., MacLachlan, V., and Healy, D. L. (1989). Predicting pregnancy outcome after in vitro fertilization and embryo transfer using estradiol, progesterone and human chorionic gonadotrophin β-subunit. Fertility and Sterility 51, 304-309.

[31] Young, D. A., Zerbe, G. O., and Hay, W. W. (1992). Application of a nonlinear random-effects model for comparing x-intercepts of linear regression lines. Statistics in Medicine 11, 2039–2040.

Article 2.4

Use of Mixed Effects Models in Cluster Analysis

Luis Villarroel del P., Guillermo Marshall R. and Anna Barón

Pontificia Universidad Católica de Chile and
University of Colorado HSC, Denver, U.S.A.

Abstract. A common situation in the biological and social sciences is to have data on one or more variables measured longitudinally on a sample of individuals. A problem of growing interest in these areas is the grouping of individuals into one of two or more clusters according to their longitudinal behavior. Recently methods have been proposed to deal with cases where individuals are classified into clusters through a linear model of mixed univariate effects deriving from a longitudinally measured variable. The method proposed in the current work deals with the case of clustering and then classification based on two or more variables measured longitudinally, through the fitting of nonlinear multivariate mixed effect models, and with consideration given to parameter estimation for balanced and unbalanced data using an EM algorithm. The application of the method is illustrated with an example in which clusters are identified and the classification into clusters is compared to the true membership of individuals in one of two groups, which is known at the end of the follow-up period.[1]

1 Introduction

It is a common situation in the course of follow-up studies in the biological and social sciences to have data for one or more variables measured in different time periods for a set of individuals. The more general case is when longitudinal information is measured over different times for each subject, neither pre-defined or equally-spaced times, and there is not complete information for all the individuals in the study. An example of this situation can be seen in the study of Anderson et al (1), based on the study of Framingham, in which the level of serum cholesterol was measured in 4,374 subjects over a period of 5 years and the effect of the change in the level of cholesterol on mortality due to cardiovascular causes was determined over 30 years of follow-up. Another example can be seen in Ellard et al (2), where the daily use of nicotine during pregnancy was measured among 338 women via a urine test, and the relationship between the amount of nicotine used and the deficit in the weight of the newborn was analyzed.

A problem of increasing interest in these areas is the grouping of individuals into two or more clusters according to their longitudinal behavior. One application of this classification could be to determine if the identified clusters coincide with some outcome of interest that occurs subsequent to the collection of the longitudinal data, which would demonstrate the utility of classification in clusters as a method of screening, or early detection, for this outcome.

Verbeke and Lesaffre (3) perform classification by clustering using linear models with mixed effects, assuming that the random effects in the model are based on a mixture of g normal

[1] Villarroel L, Marshall G, Barón AE. (2009) Cluster analysis using multivariate mixed effects models Statistics in Medicine 28 (20), 2552-2565

distributions. A SAS routine for the model of Verbeke and Lesaffre implemented with both linear and non-linear mixed models can be seen in Spiessens et al 2002 (4). Fraley and Raftery (5), propose a method of clustering based on mixture models, with applications to discriminant analysis and estimation of multivariate densities. Subsequent work of Raftery and Dean (6) addresses the problem of model selection.

Recent work, such as that of de la Cruz-Mesía, Quintana and Marshall (7) and de la Cruz-Mesía and Quintana (8) addresses clustering and discriminant analysis for longitudinal data through a method of Bayesian estimation and classification. Finally, Qin and Self (9), have proposed methodologies to deal with the case in which individuals are classified into clusters through a linear model of mixed univariate effects.

The current work proposes a method of classification into clusters, using two or more variables measured longitudinally, through the fitting of mixed non-linear multivariate models, considering the estimation of parameters for both balanced and unbalanced data. The estimation of parameters is achieved using an EM algorithm. The application of the method is illustrated with an example in which clusters are first identified, then compared in terms of classification to the true membership of each individual in one of two groups known at the end of the follow-up period.

2 Cluster Analysis using Multivariate Mixed Effects Models

To introduce the vector of multivariate responses, let Y_i be a matrix of dimension $n_i \times p$ that stores a set of p response variables for the ith subject at n_i different times. Thus, Y_i is of the form

$$Y_i = \begin{bmatrix} y_{i11} & y_{i12} & \cdots & y_{i1p} \\ y_{i21} & y_{i22} & \cdots & y_{i2p} \\ \ddots & \ddots & \ddots & \ddots \\ y_{in_i1} & y_{in_i2} & \cdots & y_{in_ip} \end{bmatrix}$$

Let E_i be the $n_i \times p$ matrix of error terms associated with Y_i. Introducing the vec operator that generates a vertical vector with the columns of Y_i, we obtain $y_i = vec(Y_i) = (y'_{i1}, y'_{i2}, \ldots, y'_{ip})'$, where y_i is the $pn_i \times 1$ vector of responses and $\epsilon_i = vec(E_i)$ is the $pn_i \times 1$ vector of errors associated with y_i.

It is assumed that the y_i follow a non-linear mixed effects model, which is described in 2.1. The clustering algorithm used to classify the individual $y_i, i = 1, \ldots, n$ is described in 2.2. Parameter estimation for the balanced case is shown in 2.3 and Appendix A, and the estimation for the unbalanced case in 2.4 and Appendix B.

2.1 The multivariate mixed effects model

It is assumed that the individual y_i belong to a fixed number of m clusters. The subject has n_i observations, that represents the maximum number of measurements for each response, but the method can acommodate missing values for individual responses. Following the notation of Lindstrom and Bates (14)), it is then assumed that y_{ijkg}, $i = 1, \ldots, n; j = 1, \ldots, n_i; k = 1, \ldots, p; g = 1, \ldots, m$, for individual i, observed at time j, on the kth response, within cluster $C_i = g$, follows a mixed non-linear model of the form

$$y_{ijkg} = f_k(\eta_{ig}, x_{ij}) + \epsilon_{ijkg} \tag{1}$$

where f_k is a non-linear function of the vector of parameters η_{ig}, the sub-index g indicates that the parameters can change across clusters $g = 1, \ldots, m$, and x_{ij} is a vector of covariates. The vector ϵ_{ijkg} represents random errors associated with y_{ijkg}. The vector of parameters η_{ig} can be incorporated into the model as $\eta_{ig} = A_i\beta_g + B_i b_{ig}$, where β_g is a vector $q \times 1$ of fixed population parameters within cluster g, b_{ig} is a $r \times 1$ vector of individual random effects that also vary according to the cluster, and the matrices A_i and B_i are design matrices of dimension $s \times q$ and $s \times r$, respectively, where s is the dimension of the vector of parameters η_{ig}.

It is assumed that $b_{ig} = b_i | C_i = g \sim MVN(0, D_g)$, with D_g of dimension $r \times r$, and $\epsilon_{ig} = \epsilon_i | C_i = g \sim MVN(0, R_{ig})$ where R_{ig} is the variance-covariance matrix of dimension $pn_i \times pn_i$. It is also assumed that, for the individual y_i and $y_{i'}$, $cov(\epsilon_{ig}, \epsilon_{i'g}) = 0$ and $cov(\epsilon_{ig}, b_{ig}) = 0$.

To give a structure to the R_{ig} matrix, it is assumed that $E_{i[j]g}$, the jth row of E_{ig}, the matrix $n_i \times p$ of random errors of Y_i within the cluster g, has distribution $E_{i[j]g} = E_{i[j]|C_i=g} \sim MVN(0, \Sigma_g)$, where Σ_g is of dimension $p \times p$, and $cov(E_{i[j]g}, E_{i[j']g}) = 0$ for all $i = 1, \ldots, n$ and $j = 1, \ldots, n_i$, except when $i = i'$ and $j = j'$. Under these assumptions, the matrix R_{ig} can be defined as

$$R_{ig} = \Sigma_g \otimes I_i \tag{2}$$

where I_i is the identity matrix of dimension $n_i \times n_i$.

Using a first-order Taylor series expansion for the initial value $\eta_{ig}^{(0)} = A_i\beta_g^{(0)} + B_i b_{ig}^{(0)}$ within the cluster g, the model (1) becomes

$$y_{ijkg} - f_k(\eta_{ig}^{(0)}, x_{ij}) + \dot{f}_k(\eta_{ig}^{(0)}, x_{ij})'\eta_{ig}^{(0)} = \dot{f}_k(\eta_{ig}^{(0)}, x_{ij})'\eta_{ig} + \xi_{ijkg} \tag{3}$$

where $\dot{f}$ is the first derivative of f with respect to η_g and ξ_{ijkg} contains both the random errors and the residuals of the function approximation. The model (3) can be written as

$$\tilde{y}_{ijkg} = \tilde{x}_{ijk}\beta_g + \tilde{z}_{ijk}b_{ig} + \xi_{ijkg} \tag{4}$$

where $\tilde{y}_{ijkg} = y_{ijkg} - f_k(\eta_{ig}^{(0)}, x_{ij}) + \dot{f}_k(\eta_{ig}^{(0)}, x_{ij})'\eta_{ig}^{(0)}$, $\tilde{x}_{ijk} = \dot{f}_k(\eta_{ig}^{(0)}, x_{ij})'A_i$ is a vector of dimension $1 \times q$, and $\tilde{z}_{ijk} = \dot{f}_k(\eta_{ig}^{(0)}, x_{ij})'B_i$ is a vector of dimension $1 \times r$. The model for the column vector with the pn_i pseudo-responses of the ith individual estimated within cluster g, is

$$\tilde{y}_{ig} = \tilde{X}_i\beta_g + \tilde{Z}_i b_{ig} + \xi_{ig} \tag{5}$$

where $\tilde{X}_i$ is a matrix of dimension $pn_i \times q$ and $\tilde{Z}_i$ is a matrix of dimension $pn_i \times r$, which is constructed with the rows of the $\tilde{x}_{ijk}$ and $\tilde{z}_{ijk}$, respectively. Note that the matrices $\tilde{X}_i$ and $\tilde{Z}_i$ are also dependent on the cluster in which they are calculated, given that both are functions of η_{ig}.

2.2 The clustering method

The method of clustering of individuals observed longitudinally follows a mixture decomposition scheme. A detailed description of the method can be seen in Theodoridis and Kostroumbas (10). Some uses of this methodology for classification can be seen in McLachlan and Gordon (11), McLachlan and Basford (12), and Qin and Self (9).

In this scheme it is assumed that there are m clusters underlying the longitudinally measured subjects, where m is a fixed number. Let C_i be a random variable with possible values 1, 2, ..., m, that indicates the cluster to which the ith subject belongs. Then, the prior probability that the subject i belongs to the gth cluster, is $P(C_i = g)$, which will be denoted by π_g. The vector $\pi = (\pi_1, \pi_2, \ldots, \pi_m)$ is unknown, and must adhere to the restrictions

$$\pi_g \geq 0 \quad g = 1, \ldots, m, \quad \sum_{g=1}^{m} \pi_g = 1 \tag{6}$$

The posterior probability that the ith subject belongs to cluster g is $P(C_i = g|y_i)$, which will be denoted by $\pi_{g|y_i}$. The classification rule is then to assign the ith subject to cluster g if $P(C_i = g|y_i) > P(C_i = j|y_i) \quad j = 1, 2, \ldots, m, \quad j \neq g.$

Assuming that C_i is not observed, we define $y_i^* = (y_i, C_i)$ as the complete data vector, where y_i is the observed part of y_i^* and C_i is the unobserved part. Assuming that the y_i^* are independent, the log-likelihood of $y^* = (y_1^*, y_2^*, \ldots, y_n^*)$ is

$$\ell(\theta) = \sum_{i=1}^{n} \sum_{g=1}^{m} \log \{P(y_i|C_i = g)P(C_i = g)\} = \sum_{i=1}^{n} \sum_{g=1}^{m} \log \{\pi_{y_i|g} \pi_g\} \tag{7}$$

where θ includes all the parameters involved in the model and $P(y_i|C_i = g)$ is the distribution of y_i, inside cluster g. This probability will be denoted by $\pi_{y_i|g}$. Then, the maximization of $\ell(\theta)$ will allow for obtaining the parameters of the assumed model for y_i in cluster g and the prior probabilities $\pi_1, \pi_2, \ldots, \pi_m$.

The EM algorithm (Dempster, Laird and Rubin (13)) is then used for the estimation of unknown parameters in the likelihood of y_i^*. Let $Q(\theta, \theta^{(t)}) = E\ell(\theta|y, \theta^{(t)})$ be the conditional expected value of the loglikelihood, conditioned on the observed part of y^* and the current parameters $\theta^{(t)}$, the parameter vector in iteration t.

The E-Step of the algorithm is

$$Q(\theta, \theta^{(t)}) = \sum_{i=1}^{n} \sum_{g=1}^{m} \pi_{g|y_i} \log \{\pi_{y_i|g} \pi_g\} \tag{8}$$

where

$$\pi_{g|y_i} = \frac{\pi_{y_i|g} \pi_g}{\sum_{g=1}^{m} \pi_{y_i|g} \pi_g} \tag{9}$$

The M-Step of the Algorithm consists of determining the parameters $\theta^{(t+1)}$ that maximize $\frac{\partial Q(\theta, \theta^{(t)})}{\partial \theta} = 0$. In this case, if the parameters in θ are functionally independent, the maximization is equivalent to

$$\sum_{i=1}^{n}\sum_{g=1}^{m}\pi_{g|y_i}\frac{\partial}{\partial\theta_g}\log\pi_{y_i|g}=0 \tag{10}$$

Using Lagrange multipliers for the restriction in (6), it is possible to obtain an expression for the prior probabilities π_g given by

$$\pi_g=\frac{1}{n}\sum_{i=1}^{n}\pi_{g|y_i} \tag{11}$$

2.3 Estimation of parameters via the EM Algorithm: Balanced Case

The distributional assumptions $b_i|C_i = g \sim MVN(0, D_g)$ and $\epsilon_i|C_i = g \sim MVN(0, R_{ig})$, with $R_{ig} = \Sigma_g \otimes I_i$, imply that $\tilde{y}_{ig} = \tilde{y}_i|C_i = g \sim MVN(\tilde{X}_i\beta_g, \tilde{Z}_iD_g\tilde{Z}_i' + \Sigma_g \otimes I_i)$. Therefore, defining $V_{ig} = \tilde{Z}_iD_g\tilde{Z}_i' + \Sigma_g \otimes I_i$ and $W_{ig} = V_{ig}^{-1}$, the probability in (9) of posterior classification into cluster g is

$$\pi_{g|\tilde{y}_i}=\frac{|V_{ig}|^{-1/2}exp\{-\frac{1}{2}(\tilde{y}_i-\tilde{X}_i\beta_g)^TW_{ig}(\tilde{y}_i-\tilde{X}_i\beta_g)\}\pi_g}{\sum_{k=1}^{m}|V_{ig}|^{-1/2}exp\{-\frac{1}{2}(\tilde{y}_i-\tilde{X}_i\beta_g)^TW_{ig}(\tilde{y}_i-\tilde{X}_i\beta_g)\}\pi_g} \tag{12}$$

where $\beta_g = \beta_g(\nu)$ and $V_{ig} = V_{ig}(\nu)$, with ν being the current iteration.

Replacing (12) in expression (10), the function to be maximized is

$$\sum_{i=1}^{n}\sum_{g=1}^{m}\pi_{g|\tilde{y}_i}\frac{\partial}{\partial\theta_g}\{-\frac{1}{2}\log|V_{ig}|-\frac{1}{2}(\tilde{y}_i-\tilde{X}_i\beta_g)^TW_{ig}(\tilde{y}_i-\tilde{X}_i\beta_g)\}=0 \tag{13}$$

subjected to restriction $\sum_{g=1}^{m}\pi_g = 1$ stated in (6).

Deriving the expression (13) with respect to β_g the maximum plausible estimator, $\tilde{\beta}_g^{(\nu)}$, is given by

$$\tilde{\beta}_g^{(\nu)}=(\sum_{i=1}^{n}\pi_{g|\tilde{y}_i}\tilde{X}_i^TW_{ig}^{(\nu)}\tilde{X}_i)^{-1}(\sum_{i=1}^{n}\pi_{g|\tilde{y}_i}\tilde{X}_i^TW_{ig}^{(\nu)}\tilde{y}_i) \tag{14}$$

For the estimation of D_g and Σ_g the EM Algorithm is used, following the methodology of estimation of variance components described by Laird and Ware (15). The complete data are $\tilde{y}_{ig}$, b_{ig}, ξ_{ig}, and C_i, the cluster to which $\tilde{y}_{ig}$ belongs. The unobserved part of the data corresponds to b_{ig}, ξ_{ig} and C_i. Successive estimations of D_g and Σ_g allow, in turn, for updating the estimations of β_g.

The sufficient statistic for Σ_g is $\sum_{i=1}^{n}E_{ig}'E_{ig}$ and the sufficient statistic for D_g is $\sum_{i=1}^{n}b_{ig}b_{ig}'$. See Appendix A for details on the sufficient statistics.

Based on these results, the first two moments of the conditional distribution of b_i given the observed data are

$$\tilde{b}_{ig}=E\{b_i|\tilde{y}_i,C_i=g,\beta_g^{(\nu)},D_g^{(\nu)},\Sigma_g^{(\nu)}\}=D_g^{(\nu)}\tilde{Z}_i'W_{ig}^{(\nu)}(\tilde{y}_i-\tilde{X}_i\beta_g^{(\nu)}) \tag{15}$$

and

$$E\{b_i b_i' | \tilde{y}_i, C_i = g, \beta_g^{(\nu)}, D_g^{(\nu)}, \Sigma_g^{(\nu)}\} = \tilde{b}_i \tilde{b}_i' + D_g^{(\nu)} - D_g^{(\nu)} \tilde{Z}_i' W_{ig}^{(\nu)} \tilde{Z}_i D_g^{(\nu)} \tag{16}$$

and the two first moments of the conditional distribution of ξ_i given the observed data are

$$\tilde{\epsilon}_{ig} = E\{\xi_i | \tilde{y}_i, C_i = g, \beta_g^{(\nu)}, D_g^{(\nu)}, \Sigma_g^{(\nu)}\} = \tilde{y}_i - \tilde{X}_i \beta_g^{(\nu)} - \tilde{Z}_i \tilde{b}_{ig} \tag{17}$$

and

$$E\{\xi_i \xi_i' | \tilde{y}_i, C_i = g, \beta_g^{(\nu)}, D_g^{(\nu)}, \Sigma_g^{(\nu)}\} = \tilde{\epsilon}_i \tilde{\epsilon}_i' + R_{ig}^{(\nu)} - R_{ig}^{(\nu)} W_{ig}^{(\nu)} R_{ig}^{(\nu)} \tag{18}$$

The calculation of the conditional expectation of the observed data constitutes the E-Step of the algorithm. The M-Step is given by

$$D_g^{(\nu+1)} = \frac{1}{n} \sum_{i=1}^{n} \pi_{g|\tilde{y}_i} \{\tilde{b}_i \tilde{b}_i' + D_g^{(\nu)} - D_g^{(\nu)} \tilde{Z}_i' W_{ig}^{(\nu)} \tilde{Z}_i D_g^{(\nu)}\} \tag{19}$$

and

$$\Sigma_g^{(\nu+1)} = \frac{1}{\sum_{i=1}^{n} \pi_{g|\tilde{y}_i} n_i} \sum_{i=1}^{n} \pi_{g|\tilde{y}_i} \{\tilde{E}_{i[j]g} \tilde{E}_{i[j]g}' + \Sigma_g^{(\nu)} - \Sigma_g^{(\nu)} W_{i[j,j]g}^{(\nu)} \Sigma_g^{(\nu)}\} \tag{20}$$

where $W_{i[j,j]g}^{(\nu)}$ is a $p \times p$ matrix with the elements of $W_{ig}^{(\nu)}$ corresponding to the observation over the time period j.

For an alternative method of estimating $\Sigma_g^{(\nu+1)}$ see Appendix A.

2.4 Estimation of parameters via the EM Algorithm: Unbalanced Case

Let O_i be a matrix with only *ones* and *zeros*, generated based on the identity matrix of dimension $pn_i \times pn_i$ (associated with the data vector y_i), from which are eliminated the rows corresponding to the missing observations. Using O_i it is possible to construct a new data vector $y_i^0 = O_i y_i$, which only contains the data actually observed for the p variables contained in the super-vector y_i. A particular case occurs when an individual doesn't have missing observations, in which case $O_i = I_{pn_i \times pn_i}$ and $y_i^0 = y_i$.

By pre-multiplying the linearized model in (5)by O_i

$$\tilde{y}_{ig}^0 = \tilde{X}_i^0 \beta_g + \tilde{Z}_i^0 b_{ig} + \xi_{ig}^0 \tag{21}$$

where $\tilde{X}_i^0 = O_i \tilde{X}_i$, $\tilde{Z}_i^0 = O_i \tilde{Z}_i$ and $\xi_{ig}^0 = O_i \xi_{ig}$. With this modification, $\tilde{y}_{ig}^0 \sim MVN(X_i^0 \beta_g, V_{ig}^0)$, with $V_{ig}^0 = \tilde{Z}_i^0 D_g^{(\nu)} Z_i^{0\prime} + O_i R_{ig}^{(\nu)} O_i'$ and R_{ig} is as defined in (2).

The estimation of the parameter β within the cluster g for the unbalanced case remains as

$$\tilde{\beta}_g^{(\nu)} = (\sum_{i=1}^{n} \pi_{g|\tilde{y}_i} \tilde{X}_i^{0'} W_{ig}^{0(\nu)} \tilde{X}_i^0)^{-1} (\sum_{i=1}^{n} \pi_{g|\tilde{y}_i} \tilde{X}_i^{0'} W_{ig}^{0(\nu)} \tilde{y}_i^0) \tag{22}$$

where $W_{ig}^{0(\nu)}$ is the inverse matrix of V_{ig}^0.

The estimation of D_g and Σ_g is

$$D_g^{(\nu+1)} = \frac{1}{n} \sum_{i=1}^{n} \pi_{g|\tilde{y}_i^0} \{\tilde{b}_i \tilde{b}_i' + D_g^{(\nu)} - D_g^{(\nu)} \tilde{Z}_i^{0'} W_{ig}^{0(\nu)} \tilde{Z}_i^0 D_g^{(\nu)}\} \tag{23}$$

and

$$\Sigma_g^{(\nu+1)} = \frac{1}{\sum_{i=1}^{n} \pi_{g|\tilde{y}_i^0} n_i} \sum_{i=1}^{n} \pi_{g|\tilde{y}_i^0} \{\tilde{E}_{i[j]g} \tilde{E}_{i[j]g}' + \Sigma_g^{(\nu)} - \Sigma_g^{(\nu)} H_{ij} O_i' W_{ig}^{0(\nu)} O_i H_{ij}' \Sigma_g^{(\nu)}\} \tag{24}$$

For further details see Appendix B.

3 An Example: Beta Sub-unit and Estradiol as Predictors of Spontaneous Abortion in Pregnancy

3.1 Models and Data

It is well known in obstetrics that, among other clinical variables, the beta-subunit of human chorionic gonadotropin (β-HCG) and estradiol shows dramatic changes in women during pregnancy. It has been established, also, that values of the subunit beta are different in women who have normal pregnancies with terminal deliveries than in women who have spontaneous abortions or other types of adverse pregnancy outcomes. This association has made it possible to classify, with some uncertainty, the outcome of pregnancy.

In a follow up study of 161 pregnant women over a period of two years in a private obstetric clinic in Santiago, Chile, the beta-subunit of human chorionic gonadotropin (β-HCG) and estradiol were measured during the first 80 days of gestation. In the study, women who had a normal pregnancy, to full term, were classified as normal, or they were classified as abnormal if they presented any complication which resulted in an interrupted pregnancy with foetal loss. Of the 161 women considered in the study, 124 had a normal delivery (77%) and 37 had an abnormal delivery (23%).

For the 161 women in the study, 348 measurements were made of β-HCG and/or estradiol; 14% of the women had one measurement, 30% had two, 44% had three and 12% had four or more measurements, with an average of 2.2 measurements per women. The missing rate was 2% for β-HCG (7 missing values) and 33.6% for estradiol (117 missing values).

For purposes of this example, the result of pregnancy (normal or abnormal delivery) was omitted from the data, with the goal of determining the predictive capacity of the clustering mechanism applied to a multivariate mixed non-linear effects model of β-HCG and estradiol as a function of gestational age.

Figure 1 shows the longitudinal trajectories of β-HCG and estradiol (log transformation) behavior over days of gestation for selected normal and abnormal subjects. A non-linear relationship is observed between log(β-HCG) and gestational age, and the logistic response curve model is a reasonable function to describe the changes in the subunit beta in the log scale across the days of pregnancy, while log(estradiol) shows a linear relationship.

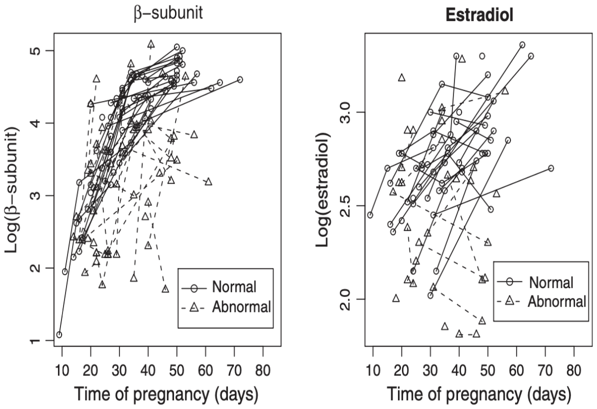


Fig. 1. Log(β-HCG) and Log(estradiol) during the first 80 days of gestation for selected normal and abnormal subjects

For the woman with id 56 from the database, the available individual information is as shown in the following matrix, which has a maximum of $n_i = 3$ observations for the $p = 2$ variables under study

$$Y_{i=56} = \begin{bmatrix} y_{i11} \; y_{i12} \\ y_{i21} \; y_{i22} \\ y_{i31} \; y_{i32} \end{bmatrix} = \begin{bmatrix} NA \; 2.15 \\ 2.90 \; NA \\ 4.19 \; 2.34 \end{bmatrix}$$

The use of the operator $y_i = vec(Y_i)$ generates a vector of the length pn_i. In the case of observation number 56, the vector is $y_{56} = (NA, 2.90, 4.19, 2.15, NA, 2.34)'$.

Letting log β-HCG be the response $k = 1$ and log estradiol be the response $k = 2$, the models for the observation $y_{ijk}|C_i = g$, with 1 random effect in each model (denoted as b_{i1g} and b_{i2g} respectively), become

$$y_{ij1g} = \frac{\beta_{1g} + b_{i1g}}{1 + \beta_{2g} \exp\{-\beta_{3g} t_{ij}\}} + \epsilon_{ij1g} \tag{25}$$

and

$$y_{ij2g} = \beta_{4g} + \beta_{5g} t_{ij} + b_{i2g} + \epsilon_{ij2g} \tag{26}$$

Details on the working variables of the logistic model can be seen in Marshall and Barón (16).

It is assumed that $b_{ig} \sim MVN(0, D_g)$ and $\epsilon_{ig} \sim MVN(0, R_{ig})$, with $R_{ig} = \Sigma_g \otimes I_{n_i}$. The models proposed in (25) and (26) are of the form $y_{ijkg} = f_g(\eta_{ig}, x_{ij}) + \epsilon_{ijkg}$ proposed in (1). To obtain a linearized version of the form $\tilde{y}_{ig} = \tilde{X}_i\beta_g + \tilde{Z}_i b_{ig} + \xi_{ig}$, the new variable response is $\tilde{y}_{ijkg} = \tilde{x}_{ijk}\beta_g + \tilde{z}_{ijk}b_{ig} + \xi_{ijkg}$, where $\beta'_g = (\beta_{1g}, \beta_{2g}, \beta_{3g}, \beta_{4g}, \beta_{5g})$ and $b'_{ig} = (b_{i1g}, b_{i2g})$.

The design matrices $\tilde{X}_i$ and $\tilde{Z}_i$ are defined are as in (3) and (4). The design matrix $\tilde{X}_i$ of dimension 2×5 ($p = 2$ response variables $\times$ $q = 5$ parameters in the vector β_g), are given by

$$\tilde{X}'_i = \begin{pmatrix} w_{ijg} & 0 \\ -f_1(\beta_g, t_{ij})exp\{\{-\beta_{3g}t_{ij}\}\}/w_{ijg} & 0 \\ f_1(\beta_g, t_{ij})\beta_{2g}t_{ij}exp\{\{-\beta_{3g}t_{ij}\}\}/w_{ijg} & 0 \\ 0 & 1 \\ 0 & t_{ij} \end{pmatrix}$$

where $w_{ijg} = 1/(1 + \beta_{2g}exp(-\beta_{3g}t_{ij}))$. The design matrix $\tilde{Z}_i$ of dimension 2×2 ($p = 2$ response variables $\times$ $r = 2$ parameters in the vector $\tilde{b}_{ig}$), is given by

$$\tilde{Z}_i = \begin{pmatrix} w_{ijg} & 0 \\ 0 & 1 \end{pmatrix}$$

Note that as the model chosen for log estradiol is linear (response $k = 2$), it fulfills $\tilde{X}_i = \beta_4 + \beta_5 t_i$ and $\tilde{Z}_i = b_{i2}$, and consequently $\tilde{y}_{i2} = y_{i2}$.

Given the structure of the incomplete data vector y_i, it is necessary to use the matrix O_i, generated on the basis of the identity matrix of dimension $pn_i \times pn_i$ associated with the vector y_i, described in the section 2.4. Then, the model for the ith observation in the gth cluster is of the form

$$\tilde{O}_i y_{ig} = O_i\tilde{X}_i\beta_g + O_i\tilde{Z}_i b_{ig} + O_i\xi_{ig} \tag{27}$$

3.2 Computational details

The algorithms used to fit the models were implemented as functions in R version 2.5 (17). Given that the parameters should be estimated within each cluster, it was necessary to obtain initial parameters for β_g, Σ_g, and D_g separately for each cluster, for which the NLME function of Pinheiro and Bates (18) was used. The estimation of parameters was performed using a tolerance of 0.005 in the log likelihood across successive iterations.

3.3 Results

To determine the maximum number of underlying clusters in the longitudinal data the Akaike Information Criterion was used (AIC) (19), defined as $AIC = -2\ell(y) + 2n_{par}$, where $\ell(y)$ is the likelihood log of the data vector y and n_{par} is the number of parameters of the fitted model. Table 1 shows the log likelihood and the value of AIC for models fitted with increasing numbers of clusters, for the models described in 3.1. Models were also fit with four or more clusters, but the high number of parameters to be estimated with the available information resulted in convergence problems in some groups, and their results were not included in the table.

Table 1. Log-likelihood and AIC for Logistic Model

Number of clusters	Log-likelihood	Number of parameters	AIC
1	270.6	10	-521.2
2	356.5	20	-673.0
3	367.8	30	-675.6

There are significant differences between the models that identify one vs. two clusters, and between the models with two vs. three clusters. Therefore, the best model fit is obtained using a model with three clusters. However, given that we know the true status of the pregnancies (normal or abnormal), we used the classification with two clusters to evaluate the ability of the model to correctly classify the women into one of the two groups.

Table 2 shows the parameters β_g estimated for the clusters $g = 1$ and $g = 2$, obtained following 32 iterations of the algorithm written in R.

Table 2. Estimated parameters

Parameter	Cluster $g = 1$	Cluster $g = 2$
β_1	4.7397	4.1830
β_2	10.2381	5.9878
β_3	-0.1493	-0.1280
β_4	2.2382	2.3337
β_5	0.0148	0.0063

The variance-covariance matrices $\hat{\Sigma}_g$ for the clusters $g = 1$ and $g = 2$, which estimate the within-subject variability in each cluster, are

$$\hat{\Sigma}_1 = \begin{pmatrix} 0.0227 & 0.0018 \\ 0.0018 & 0.0212 \end{pmatrix}, \hat{\Sigma}_2 = \begin{pmatrix} 0.2814 & 0.0439 \\ 0.0439 & 0.0446 \end{pmatrix}$$

The variance-covariance matrixes $\hat{D}_g$ for the clusters $g = 1$ and $g = 2$, which estimate the between-subject variability in each cluster, are

$$\hat{D}_1 = \begin{pmatrix} 0.0352 & 0.00018 \\ 0.00018 & 0.0265 \end{pmatrix}, \hat{D}_2 = \begin{pmatrix} 0.3624 & 0.1203 \\ 0.1203 & 0.0982 \end{pmatrix}$$

Figure 2 shows the fit of the models for log β-HCG and log estradiol in clusters $g = 1$ and $g = 2$.

Finally, through cross validation, the predictive capacity of the proposed method was evaluated for its ability to identify women with normal vs. abnormal pregnancy outcomes, heretofore assumed to be unknown. To do this, the posterior probabilities $\hat{\pi}_{g|\tilde{y}_i}$ of classification into clusters $g = 1, 2$ were estimated leaving one case out of the model at a time (leave-one-out). Using the estimated posterior probabilities for cluster 1, the area under the ROC curve (22) was calculated using as a gold standard the final pregnancy outcome. The area was equal to 0.824 with a standard error 0.042. Table 3 shows the most relevant cross-validation summary measures.

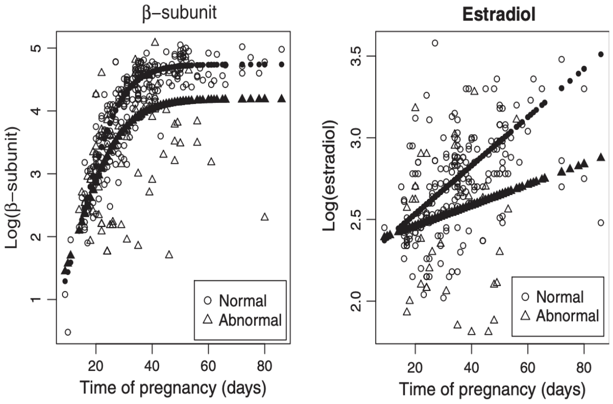


Fig. 2. Logistic model fit for Log(β-HCG) and linear model fit for Log(estradiol)

Table 3. Cross-validation of the Logistic Model

Indicator	Result
Area under ROC curve	0.824±0.042
Sensitivity (cut-off $\hat{\pi}_{1\|\tilde{y}_i} > 0.5$)	26/37 (70.3%)
Specificity (cut-off $\hat{\pi}_{1\|\tilde{y}_i} > 0.5$)	95/124 (76.6%)
Accuracy (Sens+Spec)	121/161 (75.2%)

Because we know the true outcomes for these pregnancies, we have chosen to evaluate the classification results with respect to the known outcomes, as shown in Table 3. However, an evaluation of reproducibility of the clusters would require applying other criteria, such as those indices proposed by McShane et al (24).

4 Discussion

From the point of view of the cluster analysis the methodology described in this work is a general mechanism of identification of underlying clusters using two or more longitudinally measured response variables. It is a more general methodology than previous work such as that of Verbeke and Lesaffre (3) that employs linear mixed models where the random effects arise from a mixture of distributions, or the method of Qin and Self (9), which deals with the linear univariate case. Both of these can be shown to be special cases of the framework presented here.

A limitation of the method described here is that it could be impractical if the number of clusters to be identified is very high, given that the number of conditional probabilities associated with each observation is equal to the number of components. On the other hand, given that cluster analysis is an unsupervised method of classification, there is no gold standard against which to contrast the results of the classification. Nevertheless, if the individuals are classified into one of m clusters and the estimated posterior probabilities of classification are similar to $1/m$, it would indicate that there is uncertainty in the classification. For a discussion about the uncertainty in classification via clustering see Bensmail et al 1997 (20) or Fraley and Raftery (21).

If the longitudinal data correspond to an initial step in a longer-term process, which culminates in a new response variable unknown at the outset, such as the type of birth outcome in the example above, then the classification of the individuals into clusters in accordance with their longitudinal behavior can be seen as a method of screening that could allow for identifying early those individuals who present a poor prognosis with regard to their future response.

Although, in general, classification should improve with increasing amounts of information, the use of a multivariate method requires that the correct variables are selected into the classification or prediction model. With regard to the previous point, it is interesting to note that the pregnancy outcome in the example presented above is more concordant with the posterior probability estimated with a univariate response model that considers only β-HCG rather than one which also includes estradiol. Specifically, the area under the curve in the univariate model of β-HCG is 0.83 ± 0.039, with sensitivity 78.4% and specificity 87.1%. This suggests that the selection of variables that give rise to the clusters can be critical in the resulting classification. Thus, since the clustering algorithms are unsupervised with regard to classification, it is not guaranteed that the final classification coincides with some phenomenon of interest.

Appendix A

Estimation of Parameters for the Balanced Case

Using the distribution of the complete data the conditional distribution of the sufficient statistics is obtained as

$$b_i|\tilde{y}_i, C_i = g \sim N(D_g\tilde{Z}_i'W_{ig}(\tilde{y}_i - \tilde{X}_i\beta_g), D_g - D_g\tilde{Z}_i'W_{ig}\tilde{Z}_iD_g) \tag{28}$$

and

$$\xi_i|\tilde{y}_i, C_i = g \sim N(\tilde{y}_i - \tilde{X}_i\beta_g - \tilde{Z}_ib_{ig}, R_{ig} - R_{ig}W_{ig}R_{ig}) \tag{29}$$

Based on these results, we obtain the expressions (15) and (16) for the first two moments of the conditional distribution of b_i, and the expressions (17) and (18) for the first two moments of the conditional distribution of ξ_i.

Alternatively, for the estimation of Σ_g, the conditional expectation of the rows of the E_{ig} matrix can be used, $E_{i[j]g}$. In this case,

$$\tilde{E}'_{i[j]g} = E\{E_{i[j]g}|\tilde{y}_i, C_i = g, \beta_g^{(\nu)}, D_g^{(\nu)}, \Sigma_g^{(\nu)}\} \tag{30}$$

and

$$E\{E'_{i[j]g}E_{i[j]g}|\tilde{y}_i, C_i = g, \beta_g^{(\nu)}, D_g^{(\nu)}, \Sigma_g^{(\nu)}\} = \tilde{E}_{i[j]g}\tilde{E}'_{i[j]g} + \Sigma_g^{(\nu)} - \Sigma_g^{(\nu)}W_{i[j,j]g}^{(\nu)}\Sigma_g^{(\nu)} \tag{31}$$

where $W_{i[j,j]g}^{(\nu)}$ is a $p \times p$ matrix with the elements of $W_{ig}^{(\nu)}$ corresponding to the observation over the time period j.

If $H_{ij} = I_p \otimes a_{ij}$, where a_{ij} is the jth row of the identity matrix $I_{n_i \times n_i}$, then the expression (20) for $\Sigma_g^{(\nu+1)}$ can be written as

$$\Sigma_g^{(\nu+1)} = \frac{1}{\sum_{i=1}^{n} \pi_{g|\tilde{y}_i} n_i} \sum_{i=1}^{n} \pi_{g|\tilde{y}_i} \{\tilde{E}_{i[j]g}\tilde{E}'_{i[j]g} + \Sigma_g^{(\nu)} - \Sigma_g^{(\nu)} H_{ij} W_{ig}^{(\nu)} H'_{ig} \Sigma_g^{(\nu)}\} \tag{32}$$

To illustrate the use of $H_{ij} = I_p \otimes a_{ij}$, a model with $p = 3$ response variables is assumed and an individual with $n_i = 2$ longitudinal observations. Thus, for time $j = 1$

$$H_{ij} = \begin{bmatrix} 1\,0\,0 \\ 0\,1\,0 \\ 0\,0\,1 \end{bmatrix} \otimes \begin{bmatrix} 1\,0 \end{bmatrix}' = \begin{bmatrix} 1\,0\,0 \\ 0\,0\,0 \\ 0\,1\,0 \\ 0\,0\,0 \\ 0\,0\,1 \\ 0\,0\,0 \end{bmatrix}$$

such that, for $j = 1$, the expression $H_{ij} W_{ig}^{(\nu)} H'_{ij}$ will identify the first element of the first, second and third variable explained in $W_{ig}^{(\nu)}$, in the same way in which $W_{i[j,j]g}^{(\nu)}$ does this in expression(20).

Appendix B

Estimation of Parameters for the Unbalanced Case

Using the distribution of the complete data $(\tilde{y}_i^0, C_i, \beta_g, D_g, \Sigma_g)$, the conditional distribution of the sufficient statistics for the unbalanced case,

$$b_i|\tilde{y}_i^0, C_i = g \sim N(D_g \tilde{Z}_i^{0'} W_{ig}^0 (\tilde{y}_i^0 - \tilde{X}_i^0 \beta_g, D_g - D_g \tilde{Z}_i^{0'} W_{ig}^0 \tilde{Z}_i^0 D_g) \tag{33}$$

and

$$\xi_i|\tilde{y}_i^0, C_i = g \sim N(R_{ig} O_i' W_i^0 (\tilde{y}_i^0 - \tilde{X}_i^0 \beta_g), R_{ig} - R_{ig} O_i' W_{ig}^0 O_i R_{ig}) \tag{34}$$

In a manner analogous to the balanced case, the first two moments of the conditional distribution of b_{ig} given the observed data are

$$\tilde{b}_{ig} = E\{b_i|\tilde{y}_i^0, C_i = g, \beta_g^{(\nu)}, D_g^{(\nu)}, \Sigma_g^{(\nu)}\} = D_g^{(\nu)} \tilde{Z}_i^{0'} W_{ig}^{0(\nu)} (\tilde{y}_i^0 - \tilde{X}_i^0 \beta_g^{(\nu)}) \tag{35}$$

and

$$E\{b_i b_i'|\tilde{y}_i^0, C_i = g, \beta_g^{(\nu)}, D_g^{(\nu)}, \Sigma_g^{(\nu)}\} = \tilde{b}_i \tilde{b}_i' + D_g^{(\nu)} - D_g^{(\nu)} \tilde{Z}_i' W_{ig}^{0(\nu)} \tilde{Z}_i^0 D_g^{(\nu)} \tag{36}$$

and the two first moments of the conditional distribution of ξ_i given the observed data are

$$\tilde{\epsilon}_{ig} = E\{\xi_i|\tilde{y}_i^0, C_i = g, \beta_g^{(\nu)}, D_g^{(\nu)}, \Sigma_g^{(\nu)}\} = R_{ig}^{(\nu)} O_i' W_i^{0(\nu)} (\tilde{y}_i^0 - \tilde{X}_i^0 \beta_g^{(\nu)}) \tag{37}$$

and

$$E\{\xi_i\xi_i'|\tilde{y}_i^0, C_i = g, \beta_g^{(\nu)}, D_g^{(\nu)}, \varSigma_g^{(\nu)}\} = \tilde{\epsilon}_i\tilde{\epsilon}_i' + R_{ig}^{(\nu)} - R_{ig}^{(\nu)}O_i'W_{ig}^{0(\nu)}O_iR_{ig}^{(\nu)} \tag{38}$$

If the conditional expectation of the rows of E_{ig} are considered as in (30) and (31),

$$\tilde{E}_{i[j]g}' = E\{E_{i[j]g}|\tilde{y}_i^0, C_i = g, \beta_g^{(\nu)}, D_g^{(\nu)}, \varSigma_g^{(\nu)}\} \tag{39}$$

and

$$E\{E_{i[j]g}'E_{i[j]g}|\tilde{y}_i^0, C_i = g, \beta_g^{(\nu)}, D_g^{(\nu)}, \varSigma_g^{(\nu)}\} = \tilde{E}_{i[j]g}\tilde{E}_{i[j]g}' + \varSigma_g^{(\nu)} - \varSigma_g^{(\nu)}H_{ij}O_i'W_{i[j,j]g}^{0(\nu)}O_iH_{ij}\varSigma_g^{(\nu)} \tag{40}$$

The calculation of this conditional expectation in the observed data constitutes the E-Step of the algorithm for the unbalanced case. The M-Step is given by the estimation of D_g and $\varSigma_g$ as indicated in (2.4).

Bibliography

[1] Anderson KM, Castelli WP, Levy D. Cholesterol and mortality. 30 years of follow-up from the Framingham study. *Journal of the American Medical Association* 1987; 257: 2156-2180.

[2] Ellard GA, Johnstone FD, Prescott RJ, Ji-Xian W, Jian-Hua M. Smoking during pregnancy: the dose dependence of birthweight deficits. *British Journal of Obstetrics and Gynaecology* 1996;103:806-813.

[3] Verbeke G, Lesaffre E. A linear mixed effects model with heterogeneity in the random-effects population. *Journal of the American Statistical Association* 1996; 443(91):217-221.

[4] Spiessens B, Verbeke G, Komarek A. A SAS-macro for the classification of longitudinal profiles using mixtures of normal distributions in nonlinear and generalised linear mixed model. *Technical Report* 2002.

[5] Fraley C, Raftery AE. Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association* 2002; 97:611-631.

[6] Raftery AE, Dean N. Variable selection for model-based clustering. *Technical Report no. 452* Department of Statistics, University of Washington 2004.

[7] de la Cruz-Mesía R, Quintana F, Marshall G. Model based clustering for longitudinal data. *Computational Statistics and Data Analysis* 2008; 52:1441-1457.

[8] de la Cruz-Mesía R, Quintana F. A model-based approach to Bayesian classification with applications to predicting pregnancy outcomes from longitudinal β-hCG profiles. *Biostatistics* 2007; 8(2):228-238.

[9] Qin LX, Self SG. The clustering of regression models method with applications in gene expression data. *Biometrics* 2006; 62:526-533.

[10] Theodoridis S, Kostroumbas K. Pattern recognition. Academic Press, San Diego, 1999.

[11] McLachlan GJ, Gordon RD. Mixture models for partially unclassified data: a case study of renal venous renin in hypertension. *Statistics in Medicine* 1989; 8:1291-1300.

[12] McLachlan GJ, Basford KE. Mixture models: inference and applications to clustering. Marcel Dekker, New York, 1988.

[13] Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B* 1977; 39:1-38.

[14] Lindstrom MJ, Bates DM. Nonlinear random effects models for repeated measures data. *Biometrics* 1990; 46:673-687.

[15] Laird NM, Ware JH. Random-effects models for longitudinal data. *Biometrics* 1986; 42:121-130.

[16] Marshall G., Barón AE. Linear discriminant models for unbalanced longitudinal data. *Statistics in Medicine* 2000; 19:1969-1981.

[17] R Development Core Team (2007). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

[18] Pinheiro JC, Bates DM. Mixed-effects models in S-PLUS. New York: Springer, 2000.

[19] Akaike H. A new look at the statistical identification model. *IEEE transactions on Automatic Control* 1974; 19:716-723.

[20] Bensmail H, Celeux G, Raftery AE, Robert C. Inference in model-based cluster analysis. *Statistics and Computing* 1997; 7:1-10.

[21] Fraley C, Raftery AE. How many clusters? which clustering method? answers via model-based cluster analysis. *The Computer Journal* 1998; 41:578-588.

[22] Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 1982; 143:29-36.

[23] Marshall G, De la Cruz-Mesia R, Barón AE, Rutledge J, Zerbe, GO. Nonlinear random effects model for multivariate responses with missing data. *Statistics in Medicine* 2006; 25:2817-2830.

[24] McShane LM, Radmacher MD, Freidlin B, Yu R, Li MC, Simon R. Methods for assessing reproducibility of clustering patterns observed in analyses of microarray data. *Bioinformatics* 2002; 18(11):1462-1469.

Article 2.5

Model Based Clustering for Longitudinal Data

Rolando De la Cruz-Mesía, Fernando A. Quintana, Guillermo Marshall

Pontificia Universidad Católica de Chile

Abstract. A model-based clustering method is proposed for clustering individuals on the basis of measurements taken over time. Data variability is taken into account through non-linear hierarchical models leading to a mixture of hierarchical models. We study both frequentist and Bayesian estimation procedures. From a classical viewpoint, we discuss maximum likelihood estimation of this family of models through the EM algorithm. From a Bayesian standpoint, we develop appropriate Markov chain Monte Carlo (MCMC) sampling schemes for the exploration of target posterior distribution of parameters. The methods are illustrated with the identification of hormone trajectories that are likely to lead to adverse pregnancy outcomes in a group of pregnant women.[1]

Keywords: EM-algorithm, Cluster analysis, Markov chain Monte Carlo, Mixture model, Non-linear models, Random effects.

1 Introduction

The use of mixture models for clustering is sometimes referred to as model-based probabilistic clustering (Fraley and Raftery, 1998, 2002), since a particular functional form for the component densities must be assumed. Finite mixture models are widely used for clustering data in a variety of applications (see McLachlan and Basford, 1988).

Many standard clustering algorithms are based on the assumption that the vectors to be clustered are realizations of random vectors from some parametric statistical model. These models usually place no restriction on the mean structure via covariates or otherwise. However, in many applications there is potential for parsimonious representation of the mean. For example, medical studies often yield time series-type data where each d-dimensional vector consists of measurements at d different time points. In such cases, it seems natural to model the mean via regression and we will show that there is a decided advantage in doing so, specially when tempered with the ability to detect clusters that are well defined but deviate from the model.

In longitudinal medical studies, measurements taken over time on individuals usually show a highly unbalanced structure, e.g., measurement times may be unequally spaced within a individual and may differ across individuals. Traditionally, clustering algorithms, such as K-means, have operated on points or on feature vectors of fixed-dimensional size (Hartigan, 1975). In contrast, however, data from longitudinal studies do not frequently come in a convenient fixed-dimensional form. Thus, model-based clustering has some inherent advantages compared to non-probabilistic clustering techniques (See Li, 2006). In addition to providing a generative and predictive model for the data, such methodology can conveniently handle missing and irregularly spaced measurements. Also, one key advantage of using a model-based approach is that, in addition to the

[1] De la Cruz-Mesia R, Quintana FA, Marshall G (2008) Model-based clustering for longitudinal data. Computational Statistics & Data Analysis 52 (3), 1441-1457

clustering itself, we obtain a measure of uncertainty for the assignment of each individual via the posterior probabilities of cluster membership (see (10) and (11) in Section 5). Fraley and Raftery (2002) have shown effectiveness of model-based clustering in a number of practical applications including clustering of medical data, gene expression data, web-logs data, image data, and spatial data.

In this paper, we formulate a class of regression models based on the mixture of hierarchical nonlinear models. This class can be viewed as an extension of finite mixtures of nonlinear models in which cluster specific random effects are included to account for within-cluster variability. Thus, the finite mixture of hierarchical nonlinear models provides formal estimates of individual and population probabilities of membership in each cluster, population level fixed effects for each cluster, and individual-specific random effects for each cluster conditional on membership.

Most parameter estimation methods for mixture models can be classified into two categories. One is the likelihood-based approach and the other is the Bayesian approach. Maximum likelihood estimation is greatly facilitated by the EM algorithm, while the Bayesian approach has benefited from the development of the Gibbs sampler. With the EM algorithm, latent variables (or “missing data”) are introduced, which allows finite mixture models to be fit by iteratively fitting weighted versions of the component models. So, for example, a K component finite mixture of nonlinear models can be fit via maximum likelihood (ML) by fitting K weighted nonlinear models, updating the weights and iterating to convergence. Mixture models with random effects pose an additional challenge to ML estimation as the marginal likelihood involves an integral that cannot be typically evaluated in closed form. This challenge is similar to that found with ordinary (non-mixture) hierarchical nonlinear models. Estimation in a Bayesian framework is now feasible using posterior simulation via MCMC methods. Bayes estimators for mixture models are well defined as long as the prior distributions are proper (Roeder and Wasserman, 1997). Important papers on the Bayesian analysis of mixture following MCMC methods include Diebolt and Robert (1994) and Escobar and West (1995)..

In this article, we study both frequentist and Bayesian approaches for parameter estimation. The maximum likelihood estimation is carried out via the Monte Carlo EM algorithm. From a Bayesian viewpoint, we outline a sampling strategy for fitting the models, which consists of a sequence of Gibbs and Metropolis-Hastings steps, where the latter is required when no closed form is available in some full conditional in the implementation of the Gibbs sampling algorithm.

The rest of this paper is organized as follows. In Section 2 we review related work about mixture models. We formulate the component mixture of non-linear hierarchical models in Section 3. In Section 4, we outline the EM algorithm and the Bayesian framework via MCMC methods to estimate the model. Section 5 illustrates how the methods can be used to approach the problem of clustering trajectories. In Section 6 the problem of selecting a particular mixture of hierarchical models is considered. The model class and estimation methods are illustrated with a real data example in Section 7. Finally, we give a brief discussion in Section 8.

2 Review of Clustering via Mixture Models with Regression Structure

Finite mixture models with regression structure have been extensively studied in the statistical literature and commonly applied to problems in fields such as epidemiology, medicine, genetics, economics, engineering, marketing and in the physical and social sciences. One of the earliest works was that of Quandt (1972) who defined a two-component mixture likelihood for so-called switching regressions. The methodology demonstrated the ability to find underlying group behavior by maximizing the likelihood using a conjugate gradient algorithm. Later, Quandt and Ramsey (1978) developed a procedure using the method of moments to estimate the mixture parameters for switching regressions. Hosmer (1974) also defined a two component mixture likelihood containing regression components but used maximum likelihood to estimate the mixture parameters in an iterative process. Essentially, he developed an EM algorithm for mixtures of regressions coming from two clusters. DeSarbo and Cron (1988) developed the modern EM-based procedure for mixtures of linear regressions with any number of clusters. Jones and McLachlan (1992) extend this work to multivariate data. Hurn et al. (2003) discuss solutions to the label-switching problem for Bayesian inference with regression mixtures, and Viele and Tong (2002) present consistency results of the posterior distribution.

Recently, many researchers have incorporated random effects into a wide variety of regression models to account for correlated response and multiple sources of variation. Linear and nonlinear models with fixed and random effects are two model classes that have attracted an enormous amount of attention in recent years (see Davidian and Giltinan, 1995; Vonesh and Chinchilli, 1997; Pinheiro and Bates, 2000; Verbeke and Molenberghs, 2000; Fitzmaurice et al., 2004). In a mixture model context, Gaffney and Smyth (2003) developed a random effects regression mixture framework and derived a maximum a posteriori based EM algorithm to perform inference. James and Sugar (2003) developed a functional clustering model for sparsely sampled functional data. Celeux et al. (2005) proposed a mixture of linear mixed models. They used the EM algorithm for estimating the parameters. Pfeifer (2004) considered the problem of model-based clustering based on semi-parametric mixed effects models. More recently, Booth et al. (2007) proposed a Bayesian approach to clustering multivariate data, based on a multi-level linear mixed model.

3 Mixture of Nonlinear Hierarchical Models

In this section we introduce the finite mixture of nonlinear hierarchical models and present the hormone trajectories data as a motivating application.

3.1 Model Specification

The goal of cluster analysis is to partition a collection of individuals into homogeneous subsets. Model-based clustering has recently received a great deal of attention and has provided promising results in various applications (Fraley and Raftery, 2002). In this approach, the data are viewed as coming from a mixture of distributions, each representing a different cluster. Let $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{in_i})$ denote a vector of repeated observations for individual i that is assumed to arise from one of K populations, with densities $\boldsymbol{f}_k(\boldsymbol{g}_k(\theta_{ik}, \boldsymbol{x}_{ik}); \boldsymbol{W}_{ik})$ indexed by a mean $\boldsymbol{g}_k(\cdot)$ and covariance matrix $\boldsymbol{W}_{ik}$, for $i = 1, \ldots, m$ and $k = 1, \ldots, K$. The matrix $\boldsymbol{W}_{ik}$ only depends on i for its dimension, that is, $\boldsymbol{W}_{ik}$ has dimension $n_i \times n_i$. We assume $\boldsymbol{g}_k(\cdot)$ to be a nonlinear function of unknown individual-specific parameters, θ_{ik}, and known covariates, $\boldsymbol{x}_{ik}$. For each k, the parameter vector θ_{ik}, of dimension p, follows a population distribution $\mathcal{N}_p(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$. We will assume that, conditional on $(\theta_{i1}, \ldots, \theta_{iK}, \boldsymbol{W}_{i1}, \ldots, \boldsymbol{W}_{iK}, \pi_1, \ldots, \pi_K)$, $\boldsymbol{y}_i$ follows a mixture model

$$\boldsymbol{y}_i \sim \sum_{k=1}^{K} \pi_k \boldsymbol{f}_k(\boldsymbol{g}_k(\theta_{ik}, \boldsymbol{x}_{ik}); \boldsymbol{W}_{ik}), \tag{1}$$

where π_k ($\pi_k \geq 0$, $\sum_{k=1}^{K} \pi_k = 1$) is the probability that a individual belongs to the kth cluster. Thus model (1) represents a mixture of nonlinear hierarchical models. We assume in (1) the component densities $\boldsymbol{f}_k(\boldsymbol{g}_k(\theta_{ik}, \boldsymbol{x}_{ik}); \boldsymbol{W}_{ik})$ to have multivariate normal distribution, i.e., $\boldsymbol{y}_i \sim \mathcal{N}_{n_i}(\boldsymbol{g}_k(\theta_{ik}, \boldsymbol{x}_{ik}), \boldsymbol{W}_{ik})$, if individual i belongs to cluster k. Depending on the context, various assumptions can be made about the covariance matrix $\boldsymbol{W}_{ik}$. Typically $\boldsymbol{W}_{ik}$ is required to be $\sigma_k^2 \mathbf{I}_{n_i}$, $i = 1, \ldots, m$, reflecting the assumption that individuals have exchangeable errors. In longitudinal applications, exchangeable models are common. In other situations, banded or first-order autoregressive forms where $\boldsymbol{W}_{ik}$ depends on a small number of free parameters, are more common (See De la Cruz-Mesía and Marshall, 2003, 2006 and references therein). For the remainder of this article we assume that $\boldsymbol{W}_{ik} = \sigma_k^2 \mathbf{I}_{n_i}$ in order to reduce the number of parameters to be estimated.

Each of the component densities in (1) is a individual-specific model, defined in terms of individual random-effects. The component models can be similar in form, varying only in mean specification, or have entirely different functional forms with parameters of different dimensions and meanings across submodels. A special case we use here corresponds to component models that are similar in form, and have exactly the same mean structure, but with different parameter values. Also, we assume that K has a fixed given value; discussion on how to choose K will be given in the next three sections.

A model similar to (1) was considered by Pauler and Laird (2000) who formulate a class of two-component mixtures of hierarchical nonlinear models for longitudinal data. They also outline a sampling strategy for posterior simulation, which consists of a sequence of Gibbs, Metropolis-Hastings, and reversible jump steps, where the later is required for switching between component models of different dimensions.

Logistic regression has been proposed by Hosmer and Lemeshow (2000) to classify individuals when there are known sub-populations. However, their method was not designed to handle longitudinal data. In any case, the sub-populations are not predefined in our application. Regression tree methodology, such as CART (Breiman et al., 1984), is a standard non-parametric method that can handle unknown sub-populations. However, when applying CART to longitudinal data (Segal, 1992), there are difficulties to accommodate unequally spaced and highly unbalanced data. The same problems afflict the multivariate adaptive regression splines approach for longitudinal data (MASAL) of Zhang (1997). Therefore a parametric modeling approach as we propose may provide satisfactory answers while avoiding some of the complexities inherent to more sophisticated alternatives.

3.2 Biomarker Example

Motivation for model (1) comes from a study in a private fertilization obstetrics clinic in Santiago, Chile, where the detection of pathological pregnancies was desired. Assisted reproduction treatment entails a risk of ectopic pregnancy and early pregnancy loss. Thus, early prediction of outcome is important in pregnancies following assisted reproduction treatment. In these pregnancies, the incidence of ectopic pregnancies varies from double to nearly 5-fold compared with that in spontaneous pregnancies. In particular, patients with tubal factor infertility are at an increased risk of ectopic pregnancies and should therefore receive special attention to avoid further impairment of fertility. The rate of multiple gestation is also high (20–25%) and early pregnancy loss is common, which causes anxiety in the couples involved.

Markers have been sought to distinguish between normal and abnormal pregnancies before verification of live intrauterine pregnancy by transvaginal sonography is possible. Serum Beta

Human Chorionic Gonadotropin (β-HCG) has been found to be predictive of pregnancy outcome. It is well known in obstetrics that, among other hormones, the β-HCG shows dramatic changes in women during pregnancy. It has been established, also, that values of the β-HCG are different in women who have normal pregnancies with terminal deliveries than in women who have spontaneous abortions or other types of adverse pregnancy outcomes.

In a normal pregnancy, the level of this hormone approximately doubles every 1.5 days up to 5 weeks after the last menstrual period, and then every 3.5 days from the 7th week on (Frits and Guo, 1987). After the first trimester, levels should gradually decrease over time and in fact quickly decrease to zero after the pregnancy is ended. However, abnormally large levels of β-HCG may indicate choriocarcinoma (a quick growing form of cancer that occurs in a woman's uterus after a pregnancy, miscarriage, or abortion), Down syndrome in the fetus, hydatidiform mole (a rare mass or growth that forms inside the uterus at the beginning of a pregnancy), or ovarian cancer. Lower than normal β-HCG levels may indicate ectopic pregnancy, a miscarriage or spontaneous abortion. In any case, a failure to exhibit normal growth patterns in β-HCG levels should be usually interpreted as a complicated pregnancy.

Using clinical criteria, these patients were grouped into normal and abnormal pregnancies. As reported by Marshall and Barón (2000), the normal group represents women with a normal delivery, whilst the abnormal group represents women who had any complication resulting in a non-terminal delivery and loss of the fetus.

On 173 young women, representing different pregnancies over a period of 2 years, the marker β-HCG was measured during the first 80 days of gestational age and one of the main targets of the study was to evaluate these concentrations at early stages of pregnancy, with the purpose of identifying women with a high risk of loss. Figure 1 presents the time profile in log scale for these 173 women. A non-linear relationship of the log β-HCG by day of gestational age is common for most women. We assumed these women belonged to one of $K = 2$ subpopulations: normal and abnormal pregnancies. The simple compartmental model commonly assumed for such relationship, in the kth subpopulation, leads to an underlying nonlinear model of the form

$$\boldsymbol{y}_i | z_{ik} = 1 \sim \mathcal{N}_{n_i}(\boldsymbol{g}(\theta_{ik}, \boldsymbol{t}_i); \sigma_k^2 \mathbf{I}_{n_i}), \tag{2}$$

where

$$\boldsymbol{g}(\theta_{ik}, \boldsymbol{t}_i) = \frac{\theta_{i1k}}{1 + \exp\{-(\boldsymbol{t}_i - \theta_{i2k})/\theta_{i3k}\}}.$$

Here $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{in_i})'$ denotes the longitudinal measurements on ith patient taken at arbitrary times $\boldsymbol{t}_i = (t_{i1}, \ldots, t_{in_i})'$, $i = 1, \ldots, m = 173$, and $z_{ik} = 1$ denotes that the individual i belongs to subpopulation k. We assume $\theta_{ik} = (\theta_{i1k}, \theta_{i2k}, \theta_{i3k})' \sim \mathcal{N}_3(\boldsymbol{\mu}_k, \tau_k^2 \mathbf{I}_3)$. It is easy to see that model (2) along with the population distributions above form a finite mixture of nonlinear hierarchical models (1), i.e., to analyze these data, we define a finite mixture of nonlinear hierarchical models for subpopulations of normal and abnormal pregnant women and use this to estimate individual and population probabilities of normal pregnancy.

Figure 1 also shows heterogeneous profiles. The idea is to find groups of patients with similar profiles, and ultimately link these groups to the pregnancy outcome prediction. Determining the number of groups is thus part of the inferential problem.

4 Parameter Estimation

The likelihood for model (1) is invariant under permutations of the K components. This is not a problem for a deterministic algorithm such as the EM, but it complicates the inference

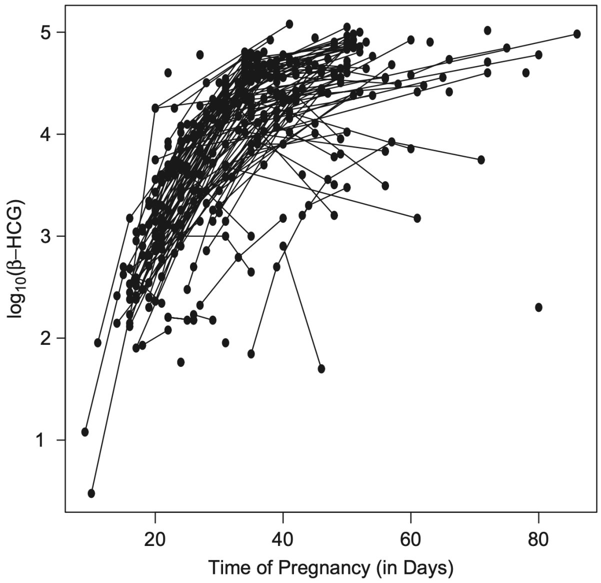


Fig. 1. Observed profiles of β-HCG for all 173 women.

from sampling procedures such as MCMC, because the labels of components may be randomly switched during the iterative process. See the discussion below.

We show first how to proceed with the EM algorithm.

4.1 MLE via an EM-type algorithm

Following McLachlan and Basford (1988), inference in mixture models is facilitated by the introduction of latent variables $\boldsymbol{z}_i = (z_{i1}, \ldots, z_{iK})$ with

$$z_{ik} = \begin{cases} 1 \text{ if } \boldsymbol{y}_i \text{ belongs to cluster } k \\ 0 \text{ otherwise.} \end{cases}$$

We assume that $\boldsymbol{z}_i$, $i = 1, \ldots, m$, are iid realizations from a multinomial distribution with probabilities $(\pi_1, \ldots, \pi_K)$ satisfying $\sum_{k=1}^K \pi_k = 1$, and that the density of $\boldsymbol{y}_i$ given $\boldsymbol{z}_i$ is

$$\prod_{k=1}^{K} [\boldsymbol{f}(\boldsymbol{g}(\theta_{ik}, \boldsymbol{x}_{ik}); \sigma_k^2 \mathbf{I}_{n_i})]^{z_{ik}}.$$

We make use of an EM-type algorithm methodology that takes into account the incomplete structure of the data. In model (1) we denote the mixture proportions by $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)$, and the nonlinear model parameters for cluster k by $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K)$, where $\boldsymbol{\beta}_k = (\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \sigma_k^2)$. Here, missing data are of two types: the indicator vectors $\boldsymbol{z} = (\boldsymbol{z}_i, i = 1, \ldots, m)$ and the random effects θ_{ik} for individual i in the kth cluster.

Then it is easy to derive the complete-data log-likelihood as

$$l(\boldsymbol{\beta}, \boldsymbol{\pi}|\boldsymbol{y}, \boldsymbol{z}, \theta) = \sum_{i=1}^{m}\sum_{k=1}^{K} z_{ik} \log(\pi_k p(\boldsymbol{y}_i, \theta_{ik}|\boldsymbol{\beta}_k))$$

where the vector $\boldsymbol{y}_i$, of size n_i, contains all the recorded values for individual i, and θ_{ik} denotes the random-effect vector for individual i in cluster k. Thus we have

$$l(\boldsymbol{\beta}, \boldsymbol{\pi}|\boldsymbol{y}, \boldsymbol{z}, \theta) = \sum_{i=1}^{m}\sum_{k=1}^{K} z_{ik} \log \pi_k + \sum_{i=1}^{m}\sum_{k=1}^{K} z_{ik} h(\boldsymbol{\beta}_k|\boldsymbol{y}_i, \theta_{ik}),$$

where

$$\begin{aligned} h(\boldsymbol{\beta}_k|\boldsymbol{y}_i, \theta_{ik}) = C &- \frac{n_i}{2}\log\sigma_k^2 - \frac{1}{2\sigma_k^2}||\boldsymbol{y}_i - \boldsymbol{g}(\theta_{ik}, \boldsymbol{x}_{ik})||^2 - \frac{1}{2}\log|\boldsymbol{\Sigma}_k| \\ &- \frac{1}{2}(\theta_{ik} - \boldsymbol{\mu}_k)'\boldsymbol{\Sigma}_k^{-1}(\theta_{ik} - \boldsymbol{\mu}_k). \end{aligned}$$

for some constant C.

E step At iteration $s > 0$, this step consists of computing the expectation of the complete log-likelihood knowing the observed data and the current value of the parameters $\boldsymbol{\beta}^{(s)}$, $\boldsymbol{\pi}^{(s)}$. In the nonlinear hierarchical model mixture context we get

$$\begin{aligned} \mathcal{Q}(\boldsymbol{\beta}, \boldsymbol{\pi}|\boldsymbol{\beta}^{(s)}, \boldsymbol{\pi}^{(s)}) &= \mathbb{E}(l(\boldsymbol{\beta}, \boldsymbol{\pi}|\boldsymbol{y}, \boldsymbol{z}, \theta)|\boldsymbol{y}, \boldsymbol{\beta}^{(s)}, \boldsymbol{\pi}^{(s)}) \\ &= \sum_{i=1}^{m}\sum_{k=1}^{K} \hat{z}_{ik}^{(s)} \log \pi_k \\ &\qquad + \sum_{i=1}^{m}\sum_{k=1}^{K} \hat{z}_{ik}^{(s)} \mathbb{E}\left[h(\boldsymbol{\beta}_k|\boldsymbol{y}_i, \theta_{ik})|\boldsymbol{y}, \boldsymbol{\beta}^{(s)}\right], \end{aligned} \tag{3}$$

where

$$\hat{z}_{ik}^{(s)} = \Pr(z_{ik} = 1|\boldsymbol{y}_i, \boldsymbol{\beta}^{(s)}, \boldsymbol{\pi}^{(s)}) = \frac{\pi_k^{(s)}\boldsymbol{p}(\boldsymbol{y}_i|\boldsymbol{\beta}_k^{(s)})}{\sum_{l=1}^{K}\pi_l^{(s)}\boldsymbol{p}(\boldsymbol{y}_i|\boldsymbol{\beta}_l^{(s)})}$$

denotes the conditional probability that $\boldsymbol{y}_i$ arises from the kth cluster, and $\boldsymbol{p}(\boldsymbol{y}_i|\boldsymbol{\beta}_k)$ is the marginal distribution obtained using Monte Carlo integration, i.e, for some large T, we compute

$$\boldsymbol{p}(\boldsymbol{y}_i|\boldsymbol{\beta}_k) \approx \frac{1}{T}\sum_{l=1}^{T}\boldsymbol{p}(\boldsymbol{y}_i|\theta_k^{(l)}, \boldsymbol{\beta}_k),$$

with $\theta_k^{(1)}, \ldots, \theta_k^{(T)} \overset{\text{iid}}{\sim} \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$.

M step This stage consists of finding the values maximizing $\mathcal{Q}(\boldsymbol{\beta}, \boldsymbol{\pi}|\boldsymbol{\beta}^{(s)}, \boldsymbol{\pi}^{(s)})$. It can now be shown that the value $(\boldsymbol{\beta}^{(s+1)}, \boldsymbol{\pi}^{(s+1)})$ of $(\boldsymbol{\beta}, \boldsymbol{\pi})$ that maximizes $\mathcal{Q}(\boldsymbol{\beta}, \boldsymbol{\pi}|\boldsymbol{\beta}^{(s)}, \boldsymbol{\pi}^{(s)})$ is given by

$$(\boldsymbol{\pi}^{(s+1)}, \boldsymbol{\mu}^{(s+1)}, \boldsymbol{\Sigma}^{(s+1)}, \sigma^{2(s+1)})'$$

where, for $k = 1, \ldots, K$

$$\begin{aligned}
\pi_k^{(s+1)} &= \frac{1}{m}\sum_{i=1}^{m} \hat{z}_{ik}^{(s)}, \\
\boldsymbol{\mu}_k^{(s+1)} &= \frac{1}{\sum_{i=1}^{m} \hat{z}_{ik}^{(s)}} \sum_{i=1}^{m} \hat{z}_{ik}^{(s)} \mathbb{E}(\theta_{ik}|\boldsymbol{y}, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \sigma_k^2), \\
\boldsymbol{\Sigma}_k^{(s+1)} &= \frac{1}{\sum_{i=1}^{m} \hat{z}_{ik}^{(s)}} \sum_{i=1}^{m} \hat{z}_{ik}^{(s)} \mathbb{E}\{(\theta_{ik} - \boldsymbol{\mu}_k^{(s+1)})(\theta_{ik} - \boldsymbol{\mu}_k^{(s+1)})'|\boldsymbol{y}, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \sigma_k^2\},
\end{aligned}$$

and

$$\sigma_k^2 = \frac{1}{\sum_{i=1}^{m} \hat{z}_{ik}^{(s)}} \sum_{i=1}^{m} \hat{z}_{ik}^{(s)} \mathbb{E}\{||\boldsymbol{y}_i - \boldsymbol{g}(\theta_{ik}, \boldsymbol{x}_{ik})||^2|\boldsymbol{y}, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \sigma_k^2\}.$$

We introduce now the following notation: let $\bar{\theta}_{ik} = \mathbb{E}(\theta_{ik}|\boldsymbol{y}, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \sigma_k^2)$, $\boldsymbol{\Theta}_{ik} = \mathrm{Cov}(\theta_{ik}|\boldsymbol{y}, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \sigma_k^2)$, $\bar{\boldsymbol{g}}_{ik} = \mathbb{E}(\boldsymbol{g}(\theta_{ik})|\boldsymbol{y}, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \sigma_k^2)$, and $\boldsymbol{\Psi}_{ik} = \mathrm{Cov}(\boldsymbol{g}(\theta_{ik})|\boldsymbol{y}, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \sigma_k^2)$. Then

$$\pi_k^{(s+1)} = \frac{1}{m}\sum_{i=1}^{m} \hat{z}_{ik}^{(s)}, \tag{4}$$

$$\boldsymbol{\mu}_k^{(s+1)} = \frac{1}{\sum_{i=1}^{m} \hat{z}_{ik}^{(s)}} \sum_{i=1}^{m} \hat{z}_{ik}^{(s)} \bar{\theta}_{ik}, \tag{5}$$

$$\boldsymbol{\Sigma}_k^{(s+1)} = \frac{1}{\sum_{i=1}^{m} \hat{z}_{ik}^{(s)}} \sum_{i=1}^{m} \hat{z}_{ik}^{(s)} \{(\bar{\theta}_{ik} - \boldsymbol{\mu}_k^{(s+1)})(\bar{\theta}_{ik} - \boldsymbol{\mu}_k^{(s+1)})' + \boldsymbol{\Theta}_{ik}\}, \tag{6}$$

and

$$\sigma_k^2 = \frac{1}{\sum_{i=1}^{m} \hat{z}_{ik}^{(s)}} \sum_{i=1}^{m} \hat{z}_{ik}^{(s)} \{||\boldsymbol{y}_i - \bar{\boldsymbol{g}}_{ik}||^2 + \mathrm{tr}(\boldsymbol{\Psi}_{ik})\}. \tag{7}$$

In the special case where $\boldsymbol{\Sigma}_k = \tau_k^2 \mathbf{I}_p$, we get

$$\tau_k^{2(s+1)} = \frac{1}{p\sum_{i=1}^{m} \hat{z}_{ik}^{(s)}} \sum_{i=1}^{m} \hat{z}_{ik}^{(s)} \{||\bar{\theta}_{ik} - \boldsymbol{\mu}_k^{(s+1)}||^2 + \mathrm{tr}(\boldsymbol{\Theta}_{ik})\}. \tag{8}$$

Thus, in order to implement the EM algorithm the quantities $\bar{\theta}_{ik}$, $\boldsymbol{\Theta}_{ik}$, $\bar{\boldsymbol{g}}_{ik}$, and $\boldsymbol{\Psi}_{ik}$ need to be evaluated at each iteration of the algorithm. One practical problem arising from our use of the EM algorithm is that there is no closed form expressions for any of $\bar{\theta}_{ik}$, $\boldsymbol{\Theta}_{ik}$, $\bar{\boldsymbol{g}}_{ik}$, or $\boldsymbol{\Psi}_{ik}$, and hence no closed form expressions for any of (5), (6) (or 8), or (7). We use Monte Carlo integration to calculate these quantities. Details are given below.

From our earlier definition $\bar{\theta}_{ik} = \int \theta_{ik}\boldsymbol{p}(\theta_{ik}|\boldsymbol{y}_i, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \sigma_k^2)\, d\theta_{ik}$. However, due the nonlinearity of random effects in the response scale, $\boldsymbol{p}(\theta_{ik}|\boldsymbol{y}_i, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \sigma_k^2)$ is not available in closed form. Nevertheless, sampling from $\boldsymbol{p}(\theta_{ik}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ is straightforward and therefore we switch to the alternative expression

$$\bar{\theta}_{ik} = \frac{\int \theta_k \boldsymbol{p}(\boldsymbol{y}_i|\theta_k, \sigma_k^2)\boldsymbol{p}(\theta_k|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\, d\theta_k}{\int \boldsymbol{p}(\boldsymbol{y}_i|\theta_k, \sigma_k^2)\boldsymbol{p}(\theta_k|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\, d\theta_k}.$$

To implement the Monte Carlo integration, take, for some large T,

$$\theta_k^{(1)}, \ldots, \theta_k^{(T)} \overset{\text{iid}}{\sim} \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

and put

$$\bar{\theta}_{ik} = \frac{\sum_{l=1}^{T} \theta_k^{(l)} \boldsymbol{p}(\boldsymbol{y}_i|\theta_k^{(l)}, \sigma_k^2)}{\sum_{l=1}^{T} \boldsymbol{p}(\boldsymbol{y}_i|\theta_k^{(l)}, \sigma_k^2)}.$$

To obtain $\boldsymbol{\Theta}_{ik}$, first put $\bar{\boldsymbol{\Theta}}_{ik} = E(\theta_{ik}\theta_{ik}'|\boldsymbol{y}_i, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \sigma_k^2)$, which is given by

$$\bar{\boldsymbol{\Theta}}_{ik} = \frac{\sum_{l=1}^{T} \theta_k^{(l)}\theta_k^{\prime(l)} \boldsymbol{p}(\boldsymbol{y}_i|\theta_k^{(l)}, \sigma_k^2)}{\sum_{l=1}^{T} \boldsymbol{p}(\boldsymbol{y}_i|\theta_k^{(l)}, \sigma_k^2)},$$

so that $\boldsymbol{\Theta}_{ik} = \bar{\boldsymbol{\Theta}}_{ik} - \bar{\theta}_{ik}\bar{\theta}_{ik}'$. Values for $\bar{\boldsymbol{g}}_{ik}$ and $\boldsymbol{\Psi}_{ik}$ are obtained in an identical manner, that is,

$$\bar{\boldsymbol{g}}_{ik} = \frac{\sum_{l=1}^{T} \boldsymbol{g}(\theta_k^{(l)}, \boldsymbol{x}_{ik})\boldsymbol{p}(\boldsymbol{y}_i|\theta_k^{(l)}, \sigma_k^2)}{\sum_{l=1}^{T} \boldsymbol{p}(\boldsymbol{y}_i|\theta_k^{(l)}, \sigma_k^2)},$$

and $\boldsymbol{\Psi}_{ik} = \bar{\boldsymbol{\Psi}}_{ik} - \bar{\boldsymbol{g}}_{ik}\bar{\boldsymbol{g}}_{ik}'$, where

$$\bar{\boldsymbol{\Psi}}_{ik} = \frac{\sum_{l=1}^{T} \boldsymbol{g}(\theta_k^{(l)}, \boldsymbol{x}_{ik})\boldsymbol{g}(\theta_k^{(l)}, \boldsymbol{x}_{ik})'\boldsymbol{p}(\boldsymbol{y}_i|\theta_k^{(l)}, \sigma_k^2)}{\sum_{l=1}^{T} \boldsymbol{p}(\boldsymbol{y}_i|\theta_k^{(l)}, \sigma_k^2)}.$$

Convergence to the true values of $\bar{\theta}_{ik}$, etc., is almost sure, so it only needs to be established which value of T leads to the required accuracy for these values.

Starting with $(\boldsymbol{\beta}^{(0)}, \boldsymbol{\pi}^{(0)})$, at sth iterative step the algorithm moves from a state $(\boldsymbol{\beta}^{(s)}, \boldsymbol{\pi}^{(s)})$ to $(\boldsymbol{\beta}^{(s+1)}, \boldsymbol{\pi}^{(s+1)})$, which is described by steps (4), (5), (6) (or (8)), and (7). Sufficient conditions for convergence are given in Dempster et al. (1977) and Wu (1983). As with all iterative searches for an MLE, a number of starting points should be considered to ensure that a true global maximum has been found.

Standard Errors A motivation for using the EM algorithm is often that the likelihood based on the observed data is difficult or impossible to evaluate. Using an EM algorithm, maximum likelihood estimates of parameters are readily obtained, but the algorithm does not immediately yield asymptotic standard errors of these estimates.

From Guo and Thompson (1994) the variance-covariance matrix $\boldsymbol{V}$ of the estimates can be written as

$$\boldsymbol{V}^{-1} = \mathbf{I}_c - \operatorname{Var}\left(\frac{\partial \log \boldsymbol{p}(\boldsymbol{y}, \boldsymbol{z}, \theta | \boldsymbol{\beta}, \boldsymbol{\pi})}{\partial(\boldsymbol{\beta}, \boldsymbol{\pi})} \middle| \boldsymbol{y} \right) \Bigg|_{(\boldsymbol{\beta},\boldsymbol{\pi})=\left(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\pi}}\right)},$$

where $\mathbf{I}_c$ is the complete data expected information matrix given by

$$\mathbf{I}_c = \mathbb{E}(\mathbf{I}_0(\boldsymbol{\beta}, \boldsymbol{\pi} | \boldsymbol{y}, \boldsymbol{z}, \theta) | \boldsymbol{y}, \boldsymbol{\beta}, \boldsymbol{\pi})|_{(\boldsymbol{\beta},\boldsymbol{\pi})=\left(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\pi}}\right)},$$

and

$$\mathbf{I}_0(\boldsymbol{\beta}, \boldsymbol{\pi} | \boldsymbol{y}, \boldsymbol{z}, \theta) = \frac{-\partial^2 \log p(\boldsymbol{y}, \boldsymbol{z}, \theta | \boldsymbol{\beta}, \boldsymbol{\pi})}{\partial(\boldsymbol{\beta}, \boldsymbol{\pi})^2},$$

where the matrix $\mathbf{I}_0(\boldsymbol{\beta}, \boldsymbol{\pi} | \boldsymbol{y}, \boldsymbol{z}, \theta)$ is given by

$$\mathbf{I}_0(\boldsymbol{\beta}, \boldsymbol{\pi} | \boldsymbol{y}, \boldsymbol{z}, \theta) = \begin{pmatrix} \mathbf{A} & 0 & 0 & 0 \\ 0 & \boldsymbol{B} & \boldsymbol{C} & 0 \\ 0 & \boldsymbol{C}' & \boldsymbol{D} & 0 \\ 0 & 0 & 0 & E \end{pmatrix}.$$

Matrices $\mathbf{A}$, $\boldsymbol{B}$, $\boldsymbol{C}$, $\boldsymbol{D}$, and E are respectively given by

$$-\frac{\partial^2 \log p(\boldsymbol{y}, \boldsymbol{z}, \theta | \boldsymbol{\beta}, \boldsymbol{\pi})}{\partial \pi_k^2}, \quad -\frac{\partial^2 \log p(\boldsymbol{y}, \boldsymbol{z}, \theta | \boldsymbol{\beta}, \boldsymbol{\pi})}{\partial \boldsymbol{\mu}_k^2}, \quad -\frac{\partial^2 \log p(\boldsymbol{y}, \boldsymbol{z}, \theta | \boldsymbol{\beta}, \boldsymbol{\pi})}{\partial \boldsymbol{\mu}_k \boldsymbol{\Sigma}_k},$$

$$-\frac{\partial^2 \log p(\boldsymbol{y}, \boldsymbol{z}, \theta | \boldsymbol{\beta}, \pi)}{\partial \boldsymbol{\Sigma}_k^2}, \quad \text{and} \quad -\frac{\partial^2 \log p(\boldsymbol{y}, \boldsymbol{z}, \theta | \boldsymbol{\beta}, \pi)}{\partial (\sigma_k^2)^2}.$$

Using formulas given in Jennrich and Schluchter (1986) we get $\mathbf{A} = \operatorname{diag}(a_1, \ldots, a_{K-1})$, with $a_k = \sum_{i=1}^{m} \left\{ \frac{z_{ik}}{\pi_k^2} + \frac{z_{iK}}{\pi_K^2} \right\}$, for $k = 1, \ldots, K-1$. Furthermore, $\boldsymbol{B} = \boldsymbol{\Sigma}_k^{-1} \sum_{i=1}^{m} z_{ik}$, and

$$C_{qr} = \boldsymbol{H}_q' \boldsymbol{\Sigma}_k^{-1} \frac{\partial \boldsymbol{\Sigma}_k}{\partial \Sigma_{kr}} \boldsymbol{\Sigma}_k^{-1} \sum_{i=1}^{m} z_{ik}(\theta_{ik} - \boldsymbol{\mu}_k)$$

for $1 \leq q \leq p$, $1 \leq r \leq p^2$, and $\boldsymbol{H}_q$ is the qth column of the $p \times p$ identity matrix $\mathbf{I}_q$. Also

$$D_{qr} = \frac{1}{2} \operatorname{tr} \left(\boldsymbol{\Sigma}_k^{-1} \frac{\partial \boldsymbol{\Sigma}_k}{\partial \Sigma_{kq}} \boldsymbol{\Sigma}_k^{-1} \left\{ \sum_{i=1}^{m} [2 z_{ik}(\theta_{ik} - \boldsymbol{\mu}_k)(\theta_{ik} - \boldsymbol{\mu}_k)' - \boldsymbol{\Sigma}_k] \right\} \boldsymbol{\Sigma}_k^{-1} \frac{\partial \boldsymbol{\Sigma}_k}{\partial \Sigma_{kr}} \right)$$

for $1 \leq q$, $r \leq p^2$, and

$$E = -\frac{1}{2\sigma_k^4} \sum_{i=1}^{m} n_i z_{ik} + \frac{1}{\sigma_k^6} \sum_{i=1}^{m} z_{ik} ||\boldsymbol{y}_i - \boldsymbol{g}(\theta_{ik}, \boldsymbol{x}_{ik})||^2.$$

In the special case $\boldsymbol{\Sigma}_k = \tau_k^2 \mathbf{I}_p$, we get

$$\begin{aligned}
\boldsymbol{B} = B &= \tau_k^{-2} \sum_{i=1}^{m} z_{ik} \\
\boldsymbol{C} = C &= \frac{1}{\tau_k^4} \sum_{i=1}^{m} z_{ik}(\theta_{ik} - \boldsymbol{\mu}_k) \\
\boldsymbol{D} = D &= -\frac{p}{2\tau_k^4} \sum_{i=1}^{m} z_{ik} + \frac{1}{\tau_k^6} \sum_{i=1}^{m} z_{ik} ||\theta_{ik} - \boldsymbol{\mu}_k||^2,
\end{aligned}$$

where, in this case, B, C and D become scalar quantities.

4.2 Bayesian Estimation via MCMC

We have seen that estimation for nonlinear hierarchical mixture models is straightforward using the EM algorithm. Estimation in a Bayesian framework can be done using posterior simulation via Markov chain Monte Carlo (MCMC) methods. Bayes estimators for mixture models are well defined so long as the prior distributions are proper. Provided that suitable (conjugate) priors are used, the posterior density will be proper, thereby allowing the application of MCMC methods such as the Gibbs sampler to provide an accurate approximation to the Bayes solution (see Roeder and Wasserman, 1997; Stephens, 2000).

Here, it is also useful to employ the latent allocation variables $\boldsymbol{z}_i$ introduced in Section 4.1. Recall the corresponding complete likelihood is

$$\prod_{i=1}^{m} \prod_{k=1}^{K} \{\pi_k p(\boldsymbol{y}_i, \theta_{ik} | \boldsymbol{\beta}_k)\}^{z_{ik}}.$$

By simplicity, the prior on $(\boldsymbol{\beta}, \boldsymbol{\pi})$ is assumed to be a product of conjugate densities

$$P(\boldsymbol{\beta}, \boldsymbol{\pi}) = P(\boldsymbol{\pi}) \prod_{k=1}^{K} P(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \sigma_k^2). \tag{9}$$

Now, we are concerned with Bayesian inference about the model parameters $\boldsymbol{\beta}$, $\boldsymbol{\pi}$ and the classification indicators $\boldsymbol{z}$. We use the Gibbs sampler for estimating parameters, as explained next.

Prior Specification We now consider the problem of choosing prior distributions for parameters $\boldsymbol{\mu}_k$, $\boldsymbol{\Sigma}_k$ and σ_k^2 of model (1). We assume prior independence for parameters in (9), i.e., $P(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \sigma_k^2) = P(\boldsymbol{\mu}_k)P(\boldsymbol{\Sigma}_k)P(\sigma_k^2)$. If the expected fraction of individuals belonging to a specific cluster is small, posterior distributions of the population and individual-specific parameters in the corresponding submodel of (1) will be poorly estimated. Therefore, proper subjective priors are necessary for all population parameters in the component densities. If subjective priors are not available, a sensitivity analysis should be performed across a range of sensible priors.

The conjugate prior on $\boldsymbol{\pi}$ will always be taken as symmetric Dirichlet distribution.

$$\boldsymbol{\pi} = (\pi_1, \dots, \pi_K) \sim \mathcal{D}(\delta, \dots, \delta),$$

and the conjugate prior distribution of $\boldsymbol{\mu}_k$, $\boldsymbol{\Sigma}_k^{-1}$, and σ_k^2 are normal, Wishart, and inverse gamma distributions, respectively:

$$\boldsymbol{\mu}_k \sim \mathcal{N}_p(\boldsymbol{\mu}_{k0}, \boldsymbol{\Sigma}_{k0}), \quad \boldsymbol{\Sigma}_k^{-1} \sim \mathcal{W}(r_{k0}, [r_{k0}\boldsymbol{R}_{k0}]^{-1}), \quad \sigma_k^2 \sim \mathcal{IG}(a_{k0}, b_{k0}).$$

The Wishart prior is parameterized such that its mean is $\boldsymbol{R}_{0k}^{-1}$. In the special case $\boldsymbol{\Sigma}_k = \tau_k^2\mathbf{I}_p$ the prior for τ_k^2 is inverse gamma, i.e.,

$$\tau_k^2 \sim \mathcal{IG}(c_{k0}, d_{k0}).$$

The inverse gamma prior is parameterized as $\pi(x) \propto x^{-(a+1)}\exp(-1/cx)$. In practice the specification of hyperparameters $\boldsymbol{\mu}_{k0}$, $\boldsymbol{\Sigma}_{k0}$, r_{k0}, $\boldsymbol{R}_{k0}$ (or c_{k0}, d_{k0}), a_{k0}, and b_{k0} may be difficult, so we can take the values of hyperparameters in such a way that we get non-informative priors in the limiting case when no (or minimal) prior information is available. For example, the prior choice for r_{k0} is $r_{k0} = p$, which is most non-informative in the sense that its distribution is flattest. Similarly, $\boldsymbol{R}_{k0}$ is chosen to be an approximate prior estimate of $\boldsymbol{\Sigma}_k$.

Full Conditionals The full conditionals for implementing Gibbs sampling are given by

$$\begin{aligned}
\boldsymbol{\pi}|\cdots &\sim \mathcal{D}(\delta + m_1, \ldots, \delta + m_K),\\
\Pr(z_{ik} = 1|\cdots) &= \frac{\pi_k^{(s)} p(\boldsymbol{y}_i|\theta_{ik}^{(s)}, \sigma_k^{2(s)})}{\sum_{l=1}^K \pi_l^{(s)} p(\boldsymbol{y}_i|\theta_{il}^{(s)}, \sigma_l^{2(s)})}, \quad k = 1, \ldots, K, \quad i = 1, \ldots, m,\\
\boldsymbol{\mu}_k|\cdots &\sim \mathcal{N}(\boldsymbol{V}_k(m\boldsymbol{\Sigma}_k^{-1}\bar{\theta}_k + \boldsymbol{\Sigma}_{k0}^{-1}\boldsymbol{\mu}_{k0}), \boldsymbol{V}_k),\\
\boldsymbol{\Sigma}_k^{-1}|\cdots &\sim \mathcal{W}(m + r_{k0}, \tilde{\boldsymbol{R}}_k),\\
\sigma_k^2|\cdots &\sim \mathcal{IG}\left(\frac{2a_{k0} + \sum_{i=1}^m n_i z_{ik}}{2}, \frac{2b_{k0}}{2 + \sum_{i=1}^m z_{ik}||\boldsymbol{y}_i - \boldsymbol{g}(\theta_{ik}, \boldsymbol{x}_{ik})||^2}\right),
\end{aligned}$$

where $\bar{\theta}_k = m^{-1}\sum_{i=1}^m z_{ik}\theta_{ik}$, $\boldsymbol{V}_k^{-1} = (m_k\boldsymbol{\Sigma}_k^{-1} + \boldsymbol{R}_{k0}^{-1})$, $\tilde{\boldsymbol{R}}_k = (r_{k0}\boldsymbol{R}_{k0} + \sum_{i=1}^m z_{ik}(\theta_{ik} - \boldsymbol{\mu}_k)(\theta_{ik} - \boldsymbol{\mu}_k)')^{-1}$, $m_k = \sum_{i=1}^m z_{ik}$, and where '$|\cdots$' denotes conditioning on all other variables. In the special case $\boldsymbol{\Sigma}_k = \tau_k^2\mathbf{I}_p$, the full conditional for τ_k^2 is

$$\tau_k^2|\cdots \sim \mathcal{IG}\left(\frac{2c_{k0} + pm_k}{2}, \frac{2d_{k0}}{2 + \sum_{i=1}^m z_{ik}(\theta_{ik} - \boldsymbol{\mu}_k)'\boldsymbol{\Sigma}_k^{-1}(\theta_{ik} - \boldsymbol{\mu}_k)}\right).$$

Generating samples from the full conditional distributions for $(\boldsymbol{z}, \boldsymbol{\pi}, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k$, (or τ_k^2), $\sigma_k^2)$ is straightforward since they all have convenient forms.

The full conditional for θ_{ik} is not available analytically. This suggests to carry out a Metropolis-Hastings step. Denote the proposal distribution by $q(\theta_{ik}|\boldsymbol{a}_{ik}, \mathbf{v}_{ik})$, which we take as a p-dimensional Normal law with mean $\boldsymbol{a}_{ik}$ and variance-covariance matrix $\mathbf{v}_{ik}$. In order to increase the efficiency of the algorithm the matrix $\mathbf{v}_{ik}$ is determined as follows. Firstly, a maximum likelihood estimate $\hat{\mathbf{v}}_{ik}$ of $\mathbf{v}_{ik}$ is obtained. Then a preliminary (random-walk) Metropolis-Hastings run is performed with the posterior distribution of θ_{ik} as the target distribution using, at this stage, $q(\theta_{ik}|\theta_{ik}^{(s)}, c_i\hat{\mathbf{v}}_{ik})$ as the proposal distribution at the $(s+1)$th iteration, where c_i is a suitable tuning parameter whose value is assumed known. After running such a chain, one obtains a sample $\{\hat{\theta}_{ik}^{(s)} : s \geq s_0\}$ from the posterior distribution of θ_{ik} and we set $\mathbf{v}_{ik} = c_i'\tilde{\mathbf{v}}_{ik}$, where $\tilde{\mathbf{v}}_{ik}$ denotes the corresponding sample variance-covariance matrix and c_i' another suitable tuning parameter chosen to get sure that the acceptance rate is satisfactory (typically between 0.2 and 0.5), see Gelman et al. (1996).

Details for implementation of MCMC algorithm are provided in the Appendix.

Label switching In finite mixture models, an identifiability problem arises from the invariance of the likelihood under permutation of the component labels unless strong prior information is used (Stephens, 2000). Under the Bayesian standpoint, this leads to symmetric and multimodal posterior distributions with up to $K!$ copies of each "genuine" mode, complicating inference on the parameters. This may cause label switching during the MCMC iterations, hence typical averages of MCMC samples of the parameters may yield unreasonable estimates of the mixture parameters. Traditional approaches to this problem impose identifiability constraints on the parameters, for instance $\pi_1 < \cdots < \pi_K$. These constraints, however, do not always solve the problem. Therefore, we adopt the relabeling procedure suggested by Celeux (1998) at each iteration of MCMC. General background on solutions that have been previously suggested for this problem can be found in Jasra et al. (2005) who categorize them to artificial identifiability constraints (Diebolt and Robert, 1994; Richardson and Green, 1997), random permutation sampling (Frühwirth-Schnatter, 2001), relabeling algorithms (Stephens, 2000; Celeux, 1998), and label invariant loss functions methods (Celeux et al., 2000; Hurn et al., 2003).

5 Allocation

From a classical viewpoint each individual i can be allocated to cluster k on the basis of the estimated posterior probabilities. Let $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\pi}})$ denote the maximum likelihood estimates of $(\boldsymbol{\beta}, \boldsymbol{\pi})$. The estimated posterior probability that individual $\boldsymbol{y}_i$ belongs to cluster k is given by

$$\boldsymbol{p}(z_{ik} = 1|\boldsymbol{y}_i, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\pi}}) = \frac{\hat{\pi}_k \boldsymbol{p}(\boldsymbol{y}_i|\hat{\boldsymbol{\beta}}_k)}{\sum\limits_{\ell=1}^{K} \hat{\pi}_\ell \boldsymbol{p}(\boldsymbol{y}_i|\hat{\boldsymbol{\beta}}_\ell)}, \tag{10}$$

where $\boldsymbol{p}(\boldsymbol{y}_i|\hat{\boldsymbol{\beta}}_k)$ is obtained via Monte Carlo integration, i.e., for some large T, we compute

$$\boldsymbol{p}(\boldsymbol{y}_i|\hat{\boldsymbol{\beta}}_k) \approx \frac{1}{T}\sum_{l=1}^{T} \boldsymbol{p}(\boldsymbol{y}_i|\theta_{ik}^{(l)}, \hat{\sigma}^2),$$

with $\theta_{ik}^{(l)} \sim \mathcal{N}_p(\hat{\boldsymbol{\mu}}_k, \hat{\boldsymbol{\Sigma}}_k)$, for $k = 1, \ldots, K$. The ith individual is then allocated according to the values of $\hat{z}_i$:

$$\hat{z}_i = \arg \max_{1 \le k \le K} \boldsymbol{p}(z_{ik} = 1|\boldsymbol{y}_i, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\pi}}),$$

i.e., to the cluster maximizing the allocation probabilities $\boldsymbol{p}(z_{ik} = 1|\boldsymbol{y}_i, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\pi}})$.

Once the label switching is taken care of, the MCMC samples can be used to draw posterior inference. Of particular interest are the allocation vector, $\boldsymbol{z}$. We compute the marginal posterior probability that individual i is allocated to cluster k as:

$$\begin{aligned}
\boldsymbol{p}(z_{ik} = 1|\boldsymbol{y}_i) &= \int \boldsymbol{p}(z_{ik} = 1|\boldsymbol{y}_i, \boldsymbol{\Omega})\boldsymbol{p}(\boldsymbol{\Omega}|\boldsymbol{y}^m)\, d\boldsymbol{\Omega} \\
&= \int \frac{\pi_k \boldsymbol{p}(\boldsymbol{y}_i|\boldsymbol{\Omega}_k)}{\sum\limits_{\ell=1}^{K} \pi_\ell \boldsymbol{p}(\boldsymbol{y}_i|\boldsymbol{\Omega}_\ell)} \boldsymbol{p}(\boldsymbol{\Omega}|\boldsymbol{y}^m)\, d\boldsymbol{\Omega} \\
&\approx \sum_{s=1}^{S} \frac{\pi_k^{(s)} \boldsymbol{p}(\boldsymbol{y}_i|\boldsymbol{\Omega}_k^{(s)})}{\sum\limits_{\ell=1}^{K} \pi_\ell^{(s)} \boldsymbol{p}(\boldsymbol{y}_i|\boldsymbol{\Omega}_\ell^{(s)})}.
\end{aligned} \tag{11}$$

where $\boldsymbol{y}^m$ denote all data and $\boldsymbol{\Omega}$ denotes all the random variables. The posterior allocation of individual i can then be estimated by the mode of its marginal posterior density:

$$\hat{z}_i = \arg \max_{1 \le k \le K} \boldsymbol{p}(z_{ik} = 1 | \boldsymbol{y}_i).$$

Class prediction

Let us now consider prediction of the class membership for a future individual $\boldsymbol{y}_f$. Using maximum likelihood, this is done by computing

$$\boldsymbol{p}(z_{fk} = 1 | \boldsymbol{y}_f, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\pi}}) = \frac{\hat{\pi}_k \boldsymbol{p}(\boldsymbol{y}_f | \hat{\boldsymbol{\beta}}_k)}{\sum_{\ell=1}^{K} \hat{\pi}_\ell \boldsymbol{p}(\boldsymbol{y}_f | \hat{\boldsymbol{\beta}}_\ell)},$$

with $\boldsymbol{p}(\boldsymbol{y}_f | \hat{\boldsymbol{\beta}}_k)$ obtained via Monte Carlo integration. Then the individual is allocated according to $\hat{z}_{fk}$:

$$\hat{z}_f = \arg \max_{1 \le k \le K} \boldsymbol{p}(z_{fk} = 1 | \boldsymbol{y}_f, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\pi}}).$$

The MCMC output can also be used to predict the class membership of future individual $\boldsymbol{y}_f$. The classification probability that a future individual $\boldsymbol{y}_f$ belongs to the kth class is

$$\begin{aligned}
\boldsymbol{p}(z_{fk} = 1 | \boldsymbol{y}_f, \boldsymbol{y}^m) &= \int \boldsymbol{p}(z_{fk} = 1 | \boldsymbol{y}_f, \boldsymbol{\Omega}) \boldsymbol{p}(\boldsymbol{\Omega} | \boldsymbol{y}^m)\, d\boldsymbol{\Omega} \\
&\propto \int \frac{\pi_k \boldsymbol{p}(\boldsymbol{y}_f | \boldsymbol{\Omega}_k)}{\sum_{\ell=1}^{K} \pi_\ell \boldsymbol{p}(\boldsymbol{y}_f | \boldsymbol{\Omega}_\ell)} \boldsymbol{p}(\boldsymbol{\Omega} | \boldsymbol{y}^m)\, d\boldsymbol{\Omega} \\
&\approx \sum_{s=1}^{S} \pi_k^{(s)} \boldsymbol{p}(\boldsymbol{y}_f | \boldsymbol{\Omega}_k^{(s)}).
\end{aligned}$$

Then the individual is allocated according to $\hat{z}_f$:

$$\hat{z}_f = \arg \max_{1 \le k \le K} \boldsymbol{p}(z_{fk} = 1 | \boldsymbol{y}_f, \boldsymbol{y}^m).$$

6 Selecting the Number of Clusters

In the previous sections, it was implicitly assumed that the number of clusters, K, was fixed. However, one of the questions of scientific interest is assessing the reliability of the output from their clustering analysis. This is equivalent to the question of determining the number of true clusters that exist in the data and for determining the number of the components in a mixture model (Roeder and Wasserman, 1997).

Several measures have been proposed for choosing the clustering model (parametrization and number of clusters); see, e.g., Chapter 6 of McLachlan and Peel (2000). We use the Bayesian Information Criterion (BIC) approximation to the Bayes factor (Schwarz, 1978) which adds a penalty to the log-likelihood based on the number of parameters, and has performed well in a number of applications (see Dasgupta and Raftery, 1998; Fraley and Raftery, 1998, 2002). The BIC has the form

$$\mathrm{BIC} = 2\widehat{\mathrm{loglik}}_{\mathcal{M}} - p_{\mathcal{M}} \log(\#\text{ of observations}), \tag{12}$$

where $\widehat{\mathrm{loglik}}_{\mathcal{M}}$ is the maximized log-likelihood for the model and data and $p_{\mathcal{M}}$ is the number of independent parameters to be estimated in the model $\mathcal{M}$. The BIC procedure is to choose the model for which the BIC criterion is maximized.

From a Bayesian viewpoint we use the Bayes factor (Kass and Raftery, 1995) as a selection tool. For the computation, we use the Chib's estimator Chib's (1995) which is based on the Gibbs output and is specially suitable for mixture models. Detailed discussion on how to compute the Bayes factor is given in Chib (1995).

7 Analysis of the Pregnant Women Data

The approaches will be illustrated by considering the clustering of cases on the basis of longitudinal measurements on pregnant women after assisted reproduction. The analysis of the β-HCG concentrations for the 173 women is carried out using model (2). Our goal is to identify clusters of trajectories and to describe any clusters that are predictive of normal or abnormal pregnancy probability.

The outcome that we analyze are the vectors of time-varying β-HCG measurements for the 173 women. Approximately 30 per cent of the 173 women had one β-HCG measurement, 31 per cent had two, 34 per cent had three, and 5 per cent had four or more measurements. The 173 women altogether contribute a total of 375 observations, where the number of samples per woman ranged from 1 through 6, with median of 2. Figure 1 presents the individual-specific $\log_{10}$ β-HCG profiles.

In order to obtain initial parameter estimates we fit a single-cluster version of the model using the NLME library of Pinheiro and Bates (2000). We next applied a model-based clustering to the random effects estimated using the MCLUST package (Fraley and Raftery, 1999) with $K = 2, 3, 4, 5$.

To facilitate fair comparison of our classical and Bayesian analyses, we selected very vague prior distributions in our analysis. That is, we use proper priors, but with hyperparameter values chosen so that the priors will have minimal impact relative to the data.

For the EM algorithm, the number T of samples for the Monte Carlo integration, was taken to be 10 000. When implementing the Gibbs sampling, we chose starting points in a neighborhood of the MLEs of model parameters. We also used other starting points obtaining similar results. We used 800 000 iterations with 100 000 sweeps as burn-in. Samples were collected at a spacing of 700 iterations, to obtain approximately independent samples. Finally we totaled 1 000 posterior Monte Carlo samples. To avoid the label switching problem, we adopt the relabeling procedure suggested by Celeux (1998) at each iteration of MCMC. To diagnose convergence, we suggest any of the convergence criteria discussed in the literature, for example, those included in the BOA package (Smith, 2004). Because of the high dimensional parameter vector, we prefer to use diagnostics, such as proposed by Geweke (1992), which do not require multiple parallel chains.

The mixture method of clustering requires the specification of the number of underlying number of clusters, K, to be fitted to the model. This decision was based on having the clinical classification of the data into two groups, but using the BIC criterion we found that the number

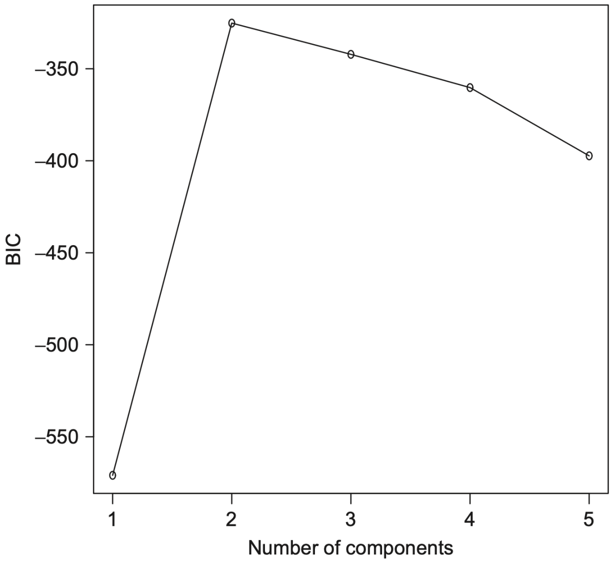


Fig. 2. Values of the BIC criterion for the data.

the cluster selected was $K = 2$ (see Figure 2). Also, we found that the Bayes factor criterion favored $K = 2$ (data not shown).

Table 1 shows parameter estimates using both methods. The results are similar. The classification results for both methods are given in Table 3. We can see that the clinical classification and the "statistical diagnosis" are different for 40 and 42 women using Bayesian and classical methods, respectively. Examination of the posterior probabilities showed that 15 of these individuals are decisively assigned to a different group than the one corresponding to the clinical classification. Although the use of Bayesian methods has only slightly improved the outright clustering, it does produce a less extreme probabilistic clustering for the misallocated women. Another comparison between the clinical classification and the model fit can be obtained by examining the estimated parameters for the model with their counterparts using the clinical classification. Agreement was fairly close.

Figure 3 shows the β-HCG trajectories for the two groups. In general, in one group we observe a steady growth of the trajectories. For the other group, however, profiles tend to have sharp increases at the start and then decrease towards the end of the window, or to have exceptionally high or low values. It is clear that the method has appropriately grouped similar women together. On the basis of maximizing classification probabilities, we see that the model classifies the normal women as being in component 1.

8 Discussion

We studied classical (EM algorithm) and Bayesian (MCMC) estimation of a proposed mixture of hierarchical nonlinear models. The model and methods can be useful in situations where repeated measures over time are available and the profiles show a nonlinear relationship across time. In

Table 1. Summary of model fitting.

	MCMC		EM	
Parameter	mean	sd	mean	sd
μ_{11}	4.749	0.037	4.762	0.038
μ_{12}	15.550	0.269	15.511	0.274
μ_{13}	6.743	0.356	6.958	0.349
μ_{21}	4.214	0.151	4.229	0.105
μ_{22}	13.970	1.129	13.923	1.145
μ_{23}	8.648	1.633	8.811	1.620
σ_1^2	0.025	0.007	0.022	0.007
σ_2^2	0.254	0.050	0.241	0.055
τ_1^2	0.026	0.015	0.032	0.012
τ_2^2	0.393	0.120	0.411	0.109
π_1	0.547	0.058	0.551	0.055

Table 2. Agreements and differences between the clinical and model classifications using Bayesian and Classical methods.

		Model classification				
classification		Bayesian		Classical		
	Groups	Normal	Abnormal	Normal	Abnormal	
	Normal	94	30	95	29	124
	Abnormal	10	39	13	36	49
		104	69	108	65	173

a cluster analysis context, this is expected to lead to more reliable clustering structures since it allows to take advantage of the powerful hierarchical nonlinear models methodology in the mixtures framework. And in many situations, it could be crucial to distinguish the statistical individuals according to their variability.

An extension of the finite mixture model that would be of particular interest in many applications is the modeling of cluster membership probabilities as a function of covariates. In the context of our example, this could be accomplished using a logistic form for the population proportion of normal pregnancies, $\pi = e^{\boldsymbol{\alpha x}}/(1+e^{\boldsymbol{\alpha x}})$ with $\boldsymbol{x}$ a vector of covariates for each women and $\boldsymbol{\alpha}$ a vector of regression parameters (Peng et al., 1996). Identifying associations between women covariates, such as age, number of previous pregnancies, and normal pregnancy tendencies can be useful for targeting specific individuals in future analysis. In our example a number of woman had missing covariate values.

Also, in our approach the number of clusters was fixed and we proposed to choose a hierarchical nonlinear model mixture with the BIC criterion. A further step we may consider, from a Bayesian viewpoint, is to treat K as random, i.e., we could formulate the clustering problem in terms of a hierarchical nonlinear model mixture with an unknown number of components and use the reversible jump Markov chain Monte Carlo technique to define a sampler that moves between different dimensional spaces. Also, in our application, the hormone trajectories are functions of the time. Functional data analysis (Ramsay and Silverman, 1997) is an ideal alternative approach for modeling such relationships. Research along these lines is currently in progress.

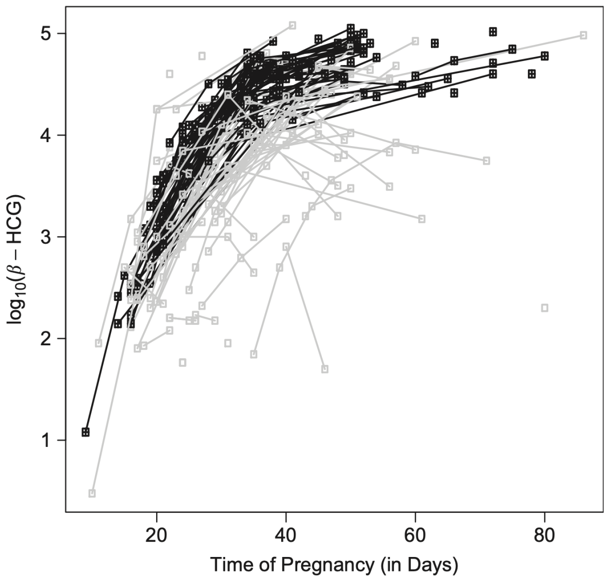


Fig. 3. Trajectories with labels.

Acknowledgments

The authors would like to thank two referees for their valuable comments which have helped to improve this paper.

Bibliography

[1] Booth, J.G., Casella, G., Hobert, J.P., 2007. Clustering using objective functions and stochastic search, submitted for publication.

[2] Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J., 1984. Classification and Regression Trees. Wadsworth, Belmont, CA.

[3] Celeux, G., 1998. Bayesian inference for mixtures: the label switching problem. In: Payne, R., Green, P. (Eds.), COMPSTAT 98. Physica-Verlag, Wurzburg, pp. 227–232.

[4] Celeux, G., Hurn, M., Robert, C.P., 2000. Computational and inferential difficulties with mixture posterior distribution. J. Amer. Statist. Assoc. 95, 957–970.

[5] Celeux, G., Lavergne, C., Martin, O., 2005. Mixture of linear mixed models for clustering gene expression profiles from repeated microarray experiments. Statist. Model. 5, 243–267.

[6] Chib, S., 1995. Marginal likelihood from the Gibbs output. J. Amer. Statist. Assoc. 90, 1313–1321.

[7] Dasgupta, A., Raftery, A.E., 1998. Detecting features in spatial point processes with clutter via model-based clustering. J. Amer. Statist. Assoc. 93, 294–302.

[8] Davidian, M., Giltinan, D.M., 1995. Nonlinear Models for Repeated Measurement Data. Chapman & Hall, London.

[9] De la Cruz-Mesía, R., Marshall, G., 2003. A Bayesian approach for nonlinear regression model with continuous errors. Comm. Statist. Theory Methods 32 (8), 1631–1646.

[10] De la Cruz-Mesía, R., Marshall, G., 2006. Nonlinear random effects models with continuous time autoregressive errors: a Bayesian approach. Statist. in Medicine 25 (9), 1471–1484.

[11] Dempster, A.P., Laird, N.M., Rubin, D.B., 1977. Maximum likelihood from incomplete data via the E-M algorithm. J. Roy. Statist. Soc. Ser. B 39, 1–38.

[12] DeSarbo, W.S., Cron, W.L., 1988. A maximum likelihood methodology for clusterwise linear regression. J. Classification 5 (1), 249–282.

[13] Efron, B., Tibshirani, R.J., 1994. Bootstrap methods for finite mixture distributions. J. Amer. Statist. Assoc. 89, 563–575.

[14] Escobar, M.D., West, M., 1994. Bayesian density estimation and inference using mixtures. J. Amer. Statist. Assoc. 90, 577–588.

[15] Fraley, C., Raftery, A.E., 1998. How many clusters? which clustering method? answers via model-based cluster analysis. Comput. J. 41, 578–588.

[16] Fraley, C., Raftery, A.E., 1999. MCLUST: software for model-based cluster analysis. J. Classification 16, 297–306.

[17] Fraley, C., Raftery, A.E., 2002. Model-based clustering, discriminant analysis and density estimation. J. Amer. Statist. Assoc. 97, 611–631.

[18] Frits, M.A.O., Guo, S.M., 1997. Doubling time of human chorionic gonadotropin (hCG) in early normal pregnancy: relationship to hCG concentration and gestational age. Fertil. Steril. 47, 584–589.

[19] Frühwirth-Schnatter, S., 2001. Markov chain Monte Carlo estimation of classical and dynamic switching and mixture models. J. Amer. Statist. Assoc. 96, 194–209.

[20] Gaffney, S.J., Smyth, P., 2003. Curve clustering with random effects regression mixtures. In: Bishop, C.M., Frey, B.J. (Eds.), Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics. Key West, FL.

[21] Gelman, A., Roberts, G.O., Gilks, W.R., 1995. Efficient Metropolis jumping rules. In: Bernardo, J.M., Berger, J.O., Dawid, A.P., Smith, A.F.M. (Eds.), Bayesian Statistics, vol. 5. Oxford University Press, Oxford, pp. 599–607.

[22] Geweke, J., 1992. Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. In: Bernardo, J.M., Berger, J.O., Dawid, A.P., Smith, A.F.M. (Eds.), Bayesian Statistics, vol. 4. Oxford University Press, Oxford, pp. 169–194.

[23] Guo, S.W., Thompson, E.A., 1994. Monte Carlo estimation of mixed models for large complex pedigrees. Biometrics 50, 417–432.

[24] Hartigan, J.A., 1975. Clustering Algorithms. Wiley, New York.

[25] Hosmer, D.W., Lemeshow, S., 2000. Applied Logistic Regression, second ed. Wiley, New York.

[26] Hurn, M., Justel, A., Robert, C.P., 2003. Estimating mixture of regressions. J. Comput. and Graphical Statist. 12 (1), 55–79.

[27] Ishwaran, H., James, L.F., 2002. Gibbs sampling methods for stick breaking priors. J. Amer. Statist. Assoc. 97, 161–179.

[28] Jasra, A., Holmes, C.C., Stephens, D.A., 2005. Markov chain Monte Carlo and the label switching problem in Bayesian mixture modelling. Statist. Sci. 20 (1), 50–67.

[29] Jennrich, R.I., Schluchter, M.D., 1986. Unbalanced repeated-measures models with structured covariance matrices. Biometrics 42 (4), 805–820.

[30] Jones, P.N., McLachlan, G.J., 1992. Fitting finite mixture models in a regression context. Austral. J. Statist. 34 (2), 233–240.

[31] Kass, R.E., Raftery, A.E., 1995. Bayes factors. J. Amer. Statist. Assoc. 90, 773–795.

[32] Li, B., 2006. A new approach to cluster analysis: the clustering-function-based method. J. Roy. Statist. Soc. Ser. B 68, 457–476.

[33] Marsh, G., Barón, A.E., 2000. Linear discriminant models for unbalanced longitudinal data. Statist. in Medicine 19, 1969–1981.

[34] McLachlan, G.J., Basford, K.E., 1988. Mixture Models: Inference and Applications to Clustering. Marcel Dekker, New York.

[35] McLachlan, G.J., Peel, D., 2000. Finite Mixture Models. Wiley, New York.

[36] Müller, D.K., Stadtmüller, U., 2000. A mixture model for longitudinal data with application to assessment of noncompliance. Biometrics 56, 464–472.

[37] Peng, F., Jacobs, R.A., Tanner, M.A., 1996. Bayesian inference in mixtures-of-experts and hierarchical mixtures-of-experts with an application to speech recognition. J. Amer. Statist. Assoc. 91, 953–960.

[38] Pinheiro, J.C., Bates, D.M., 2000. Mixed-Effects Models in S and S-PLUS. Springer, New York.

[39] Quandt, R.E., 1972. A new approach to estimating switching regressions. J. Amer. Statist. Assoc. 57, 306–310.

[40] Randolph, R.E., Ramsey, J.O., 1978. Estimating mixtures of normal distributions and switching regressions. J. Amer. Statist. Assoc. 73, 730–738.

[41] Ramsay, J.O., Silverman, B.W., 1997. Functional Data Analysis. Springer, New York.

[42] Richardson, S., Green, P.J., 1997. On Bayesian analysis of mixture models with an unknown number of components. J. Roy. Statist. Soc. Ser. B 59 (4), 731–792.

[43] Roeder, K., Wasserman, L., 1997. Practical Bayesian density estimation using mixtures of normals. J. Amer. Statist. Assoc. 92, 894–902.

[44] Schwarz, G., 1978. Estimating the dimension of a model. Ann. Statist. 6, 461–464.

[45] Seber, G.A.F., Wild, C.J., 1989. Nonlinear Regression. Wiley, New York.

[46] Smith, B.J., 2004. Bayesian Output Analysis Program (BOA). Version 1.1.2 for S-Plus and R. Available at: `http://www.public-health.uiowa.edu/boa/`.

[47] Stephens, M., 2000. Dealing with label switching in mixture models. J. Roy. Statist. Soc. Ser. B 62, 795–809.

[48] Verbeke, G., Molenberghs, G., 2000. Linear Mixed Models for Longitudinal Data. Springer, New York.

[49] Wang, Y., Tang, B., 2002. Modelling nonlinear mixed effects with linear regressions. Ann. Statist. 27, 439–460.

[50] West, M., Müller, P., Escobar, M.D., 1994. Hierarchical Priors for Mixture Models. Technical Report 94-A02, Institute of Statistics and Decision Sciences, Duke University.

[51] Wu, C.F.J., 1983. On the convergence properties of the EM algorithm. Ann. Statist. 11 (1), 95–103.

[52] Zhang, H., 1997. Multivariate adaptive splines for analysis of longitudinal data.

Appendix

MCMC Algorithm Implementation Details, with superscripts indicating the iteration.

1. $s = 0$. Fix initial values $\boldsymbol{z}^{(0)}, \boldsymbol{\pi}^{(0)}, \boldsymbol{\mu}^{(0)}, \boldsymbol{\Sigma}^{(0)}(\text{or } \tau^{2(0)}), \sigma^{2(0)}, \theta^{(0)}$.
2. Sample $z_i^{(s+1)}$ from
$$p_{ik} = \frac{\pi_k^{(s)} p(\boldsymbol{y}_i | \theta_{ik}^{(s)}, \sigma_k^{2(s)})}{\sum_{l=1}^K \pi_l^{(s)} p(\boldsymbol{y}_i | \theta_{il}^{(s)}, \sigma_l^{2(s)})}, \quad k = 1, \dots, K,\ i = 1, \dots, m.$$
3. Sample $\boldsymbol{\pi}^{(s+1)} = (\pi_1^{(s+1)}, \dots, \pi_K^{(s+1)})$ from $\mathcal{D}(\delta + m_1^{(s+1)}, \dots, \delta + m_K^{(s+1)})$ where $m_k^{(s+1)} = \sum_{i=1}^m z_{ik}^{(s+1)}$.
4. Sample $\boldsymbol{\mu}_k^{(s+1)}$ from
$$\mathcal{N}\left(\boldsymbol{V}_k(m\boldsymbol{\Sigma}_k^{-1}\bar{\theta}_k + \boldsymbol{\Sigma}_{k0}^{-1}\boldsymbol{\mu}_{k0}), \boldsymbol{V}_k\right).$$
5. Sample $\boldsymbol{\Sigma}_k^{(s+1)}$ from
$$\mathcal{W}(m + r_{k0}, \tilde{\boldsymbol{R}}_k).$$
   5.1 In the special case $\boldsymbol{\Sigma}_k = \tau_k^2 \mathbf{I}_p$, sample $\tau_k^{2(s+1)}$ from
$$\mathcal{IG}\left(\frac{2c_{k0} + pm_k}{2}, \frac{2d_{k0}}{2 + \sum_{i=1}^m z_{ik}(\theta_{ik} - \boldsymbol{\mu}_k)'\boldsymbol{\Sigma}_k^{-1}(\theta_{ik} - \boldsymbol{\mu}_k)}\right).$$
6. Sample $\sigma_k^{2(s+1)}$ from
$$\mathcal{IG}\left(\frac{2a_{k0} + \sum_{i=1}^m n_i z_{ik}}{2}, \frac{2b_{k0}}{2 + \sum_{i=1}^m z_{ik}||\boldsymbol{y}_i - \boldsymbol{g}(\theta_{ik}, \boldsymbol{x}_{ik})||^2}\right).$$
7. Sample from the full conditional distribution of $\theta_{ik}^{(s+1)}$ given the observations $\boldsymbol{y}$ and the (vector of) parameters
$$(\boldsymbol{z}^{(s+1)}, \boldsymbol{\pi}^{(s+1)}, \boldsymbol{\mu}^{(s+1)}, \boldsymbol{\Sigma}^{(s+1)}(\text{or } \tau^{2(s+1)}), \sigma^{2(s+1)})$$
   via a (random walk) Metropolis-Hastings step with proposal distribution
$$q(\theta_{ik}^{(s+1)} | \theta_{ik}^{(s)}, \mathbf{v}_{ik}).$$
8. $s = s + 1$. Go to step 2.

Table 3. Agreements and differences between the clinical and model classifications using Bayesian and Classical methods

Clinical classification		Model classification								
		RE				No RE				
		Bayesian		Classical		Bayesian		Classical		
	Groups	N	A	N	A	N	A	N	A	
	Normal(N)	94	30	95	29	93	31	99	25	124
	Abnormal(A)	10	39	13	36	12	37	29	40	49
		104	69	108	65	105	68	128	45	173

Chapter 3

Arsenic Exposure and Human Health

Article 3.1

Fifty-Year Study of Lung and Bladder Cancer Mortality in Chile Related to Arsenic in Drinking Water

Guillermo Marshall, Catterina Ferreccio, Yan Yuan, Michael N. Bates, Craig Steinmaus, Steve Selvin, Jane Liaw, Allan H. Smith

Universidad Católica de Chile and School of Public Health, University of California, Berkeley, CA

Abstract. Backgroud Region II of Chile (the second most northerly administrative region) experienced dramatic increases in average arsenic water concentrations beginning in 1958, followed by marked declines in the 1970s when water treatment plants were installed. This history provides a unique opportunity to study time trends in the development of arsenic-related cancers, including lung and bladder cancers.

Methods We investigated lung and bladder cancer mortality from 1950 to 2000 for region II compared with region V, where drinking water was not contaminated with arsenic. Mortality data were obtained from 218174 death certificates for the two regions for 1950?1970 and from mortality data tapes that identified 307541 deaths in the two regions for 1971?2000. Poisson regression models were used to identify time trends in rate ratios (RRs) of mortality from lung and bladder cancers comparing region II

Results Lung and bladder cancer mortality rate ratios for region II compared with region V started to increase about 10 years after high arsenic exposures commenced and continued to rise until peaking in 1986?1997. The peak lung cancer mortality RRs were 3.61 (95% confidence interval [CI] = 3.13 to 4.16) for men and 3.26 (95% CI = 2.50 to 4.23) for women. The peak bladder cancer RRs were 6.10 (95% CI = 3.97 to 9.39) for men and 13.8 (95% CI = 7.74 to 24.5) for women. Combined lung and bladder cancer mortality rates in region II were highest in the period 1992?1994, with mortality rates of 153 and 50 per 100000 men and women, respectively, in region II compared with 54 and 19 per 100000 in region V.

Conclusions Such large increases in total population cancer mortality rates have, to our knowledge, not been docu- mented for any other environmental exposure. The long latency pattern is noteworthy, with mortality from lung and bladder cancers continuing to be high until the late 1990s, even though major decreases in arsenic exposure occurred more than 25 years earlier. [1]

Drinking water in region II of Chile is supplied mainly by rivers originating from springs in the Andes mountains that contain inorganic arsenic, some at very high concentrations. Before 1958, the arsenic concentration in the water supply in the main city of region II, Antofagasta, was approximately 90 µg/L (1), nearly twice the drinking water standard in much of the world (50 µg/L) until the recent lowering of the level in some countries (to 10 µg/L in most cases). In 1958, this source was supplemented with water from the Toconce and Holajar rivers, which contain extremely high levels of arsenic (1),; as a result, the arsenic concentrations in the water supply

[1] Marshall G, Ferreccio C, Yuan Y, et al. Fifty-year study of lung and bladder cancer mortality in Chile related to arsenic in drinking water. *J Natl Cancer Inst* 2007; 99: 920–8.

of Antofagasta and the nearby city of Mejillones averaged 870 µg/L during 1958 – 1970. Other cities and towns in the region also had high concentrations of arsenic in their water supplies. Installation of a water treatment plant in Antofagasta in 1971, and later in other cities and towns, led to marked reductions in arsenic concentrations in the water.

The sudden rise and subsequent fall of arsenic concentrations in region II provides a unique opportunity to investigate time trends in arsenic-associated cancer. Also, by contrast to other high-arsenic areas in the world, where most of the arsenic is obtained from well water and historical exposure reconstruction is difficult, Northern Chile has a very dry climate, and each town and city has just one or two sources of water originating from rivers for which arsenic measurements have been recorded since the 1950s. The arsenic monitoring data for region II water supplies provide exposure data of a quality far exceeding that of other arsenic- exposed regions of the world, for which past exposures are invariably difficult to ascertain.

Increased mortality from lung and bladder cancers has previously been reported in region II of Chile compared with the rest of the country. These cancers have also been associated with high levels of arsenic in water supplies in Taiwan (2; 3; 4; 5) and Argentina (6; 7). The results from Chile confirmed that the elevated cancer rates in these other countries were likely to be attributable to arsenic, and in 2002, the Working Group of the International Agency for Research on Cancer classified arsenic in drinking water as a cause of lung and bladder cancers, along with skin cancer (8).

Little is known about the latency period from commencement of increased exposure to arsenic to increased risk of cancer. Some studies have suggested that latency periods may be more than 30 years long (9; 10; 11; 12; 13; 14). The relatively sharp peak in water arsenic concentrations in region II of Chile during 1958 – 1970 allows construction of longitudinal mortality time patterns that can be used to investigate latency periods associated with diseases caused by arsenic exposure, including cancer.

We previously reported marked increases in lung and bladder cancer mortality in the years 1989 – 1993 in arsenic-exposed region II of Chile compared with the rest of Chile (15). These findings led us to investigate mortality in region II for the 50-year period 1950-2000. In this paper, we present lung and bladder cancer mortality for these years in region II compared with region V, which is otherwise similar to region II but not exposed to arsenic in drinking water. The unique exposure scenario—in a large population with well-documented information on past exposure via drinking water—provides a rare opportunity to investigate the latency effects of a widespread environmental carcinogen, including the latency period between the reduction in exposure to reductions in cancer rates.

1 Subjects and Methods

Setting

Chile is a long, narrow country that is divided into 15 administrative regions that are numbered from north to south, with region II being the second northernmost (Fig. 1). A single comparison region for region II was chosen because International Classification of Diseases (ICD) coding of death certificates for the whole country for a period of 20 years (from 1950 to 1970, when computerized data were not already available) would have been prohibitively expensive. Because the country varies in characteristics from north to south, in particular in factors related to climate, we wished to select a nearby region for comparison. To increase statistical precision, it was also desirable that the number of people in the referent population be greater than the number in region II. Regions I and III were first considered, but both were rejected because they are adjacent to region II and there was an increased potential for migration between them and because there has been some arsenic expo- sure in these two regions, although it is minor compared with that in region II (16). Region IV is small, with population numbers similar to those in region II. Region V is located in the northern half of Chile, with a population more than three times larger than that of region II. In 2000, the population of region II was 477.332 and that of region V was 1.508.749; the ratio has been similar throughout the study period.

To ensure that region V was an appropriate choice of comparison region, in preliminary work (data not shown) we compared key sociodemographic factors and relevant medical information among region II, region V, and the whole of Chile. Per capita income in region V in 1990 (US $2053) was similar to that of the country as a whole (US $2011) (17). Region II had a much higher per capita income (US $3853), but the difference mainly reflected the presence of the mining industry in region II, which generated large exports but not higher personal income. Data from smoking surveys carried out on random population samples in 1990 and 1992 (18) suggest little in the way of smoking differences between the two regions and the rest of the country, with both years giving similar data. In 1990, 26.6% of surveyed men and 19.3% of surveyed women in Chile said they smoked. The corresponding figures for region II and region V were, respectively, 27.4% and 28.5% for men and 16.6% and 20.2% for women. Finally, we analyzed information concerning death certification by regions in the country, based on a study conducted in 1983 (19). For the whole country, 85.6% of the death certificates in that year were certified by physicians, and region II and region V had similar percentages, with 89.8% and 94.5% physician certification, respectively.

Mortality Data

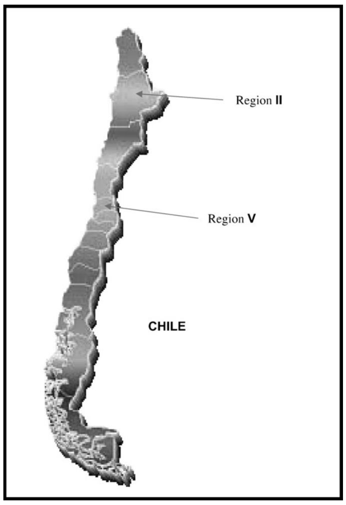
Mortality data covering the period 1950 − 2000 for individuals aged 30 years and above were obtained from three different sources. For the period 1950 − 1970, a period for which electronic mortality data were not available, we obtained death certificates for regions II and V from the Chilean Civil Registry and Identification Department. Digital photographs were taken of the 218 174 death certificates for regions II and V and they were downloaded to a central computer, after which study nosologists coded the causes of death according to the International Classification of Diseases, 9th Revision (20). The nosologists were given a mixture of death certificates from each region and coded causes of death without knowing from which region the death certificate originated. A separate person entered the data on the region from which the death certificates originated.

For the periods 1971–1975 and 1977–1982, computerized mortality data that included cause of death were obtained from the Chile National Institute of Statistics (Instituto Nacional de Estadísticas) including 52 155 deaths for regions II and V for 1971–1975, and 58 638 deaths for 1977 − 1982. However, no mortality data are available for the country for the year 1976. Finally, for the period 1983–2000, mortality data were obtained from the Ministry of Health for all of Chile including 196 748 deaths for regions II and V.

Census data were used to calculate the denominators for mortality rates. Chile has had a total population census roughly every 10 years, including 1940, 1952, 1960, 1970, 1982, 1992, and 2002. Census data were obtained for region II, region V, and the rest of Chile from the National Institute of Statistics (Instituto Nacional de Estadísticas) for men and women separately in 10-

year age groups. We estimated population counts with a linear interpolation for years between each census.

Fig. 1. Map of Chile, showing regions II and V. The country is adminis- tratively divided into regions that are numbered from north to south.



Data on Arsenic Levels in Drinking Water

Data on water arsenic concentrations for cities and towns in region II from 1950 to 1994 were obtained from a previous study (15), in which approximate average levels of arsenic in sources of drinking water were given for all towns and cities in region II.

Statistical Methods

Lung cancer (ICD code 162) and bladder cancer (ICD code 188) mortality rate ratios for region II compared with region V were estimated using Poisson regression for each 3 years of calendar time between 1950 and 2000, for men and women separately. Because these cancers are rare in persons below age 30 years, the analysis was restricted to those aged 30 years and above. Age adjustment incorporated 10-year age groups starting with ages 30 – 39 years and continuing to ages 80 years and above. Ten-year age groups were chosen because the census data were available in that form. From 1971 through 2000, it was possible to estimate rate ratios for region II compared with all of the rest of Chile; 95% confidence intervals (CIs) were calculated based on the asymptotic normality of the logarithm of the rate ratio estimate (21). Poisson regression models were used to smooth the effect of arsenic exposure over time. In these models, the response variable was the number of observed deaths in each age group for each year, for both region II and region V. The model for the expected number of deaths in age group i , region j , and year t can be expressed as

$$\log \mu_{ijt} = \alpha + \text{age}_i + \text{region}_j + \text{year}_t + f(t, \text{region}_j) + \log \text{population}_{ijt},$$

where $f(t, \mathrm{region}_j)$ is a cubic spline representing the interaction between year and region, with region V as reference. The log rate ratio for period t between regions II and V can be represented as

$$\log \mathrm{RR}_t = \mathrm{region}_2 + f(t, \mathrm{region}_2)$$

This model for analyzing time trends was first presented by Hastie and Tibshirani (22) and has subsequently been used with Poisson regression by many authors (23; 24). We used spline smoothing with generalized cross-validation to estimate the amount of smoothing. We also plotted mortality rate ratios estimated for each successive 3-year period, along with the smoothed function and its 95% confidence interval, to check that the smoothed function was compatible with the underlying mortality data. Finally, we separated the mortality data into three 20-year birth cohorts chosen in relation to 1958, when the high arsenic exposures commenced, and repeated the Poisson regression analysis for each birth cohort separately. The birth cohorts were chosen to identify childhood exposure (those born in 1938–1957) and two earlier birth cohorts, one having young adult exposure who were born in the period 1918–1937 and the second with older age adult exposure who were born before 1918.

2 Results

Water arsenic concentrations for towns and cities in region II, and the overall population-weighted averages, are presented in Table 1 , which was published previously (15). To place these arsenic concentrations in water into context, the highest population-weighted average for region II was 569 µ g/L, a little more than 10 times the 50 µ g/L level that was, until recently, the drinking water standard in much of the world. Arsenic concentrations in water are low in the rest of Chile, including region V. For example, the water arsenic concentrations in tap water in Valparaíso, the largest city of region V, were in the range of 0.5 − 1.1 µ g/L when sampled in 1998 (25).

We calculated lung cancer mortality rates and mortality rate ratios (RRs) for men and women separately, comparing region II with region V for the period 1950–2000 and comparing region II with the rest of Chile for the period 1971–2000 (Table 2). The peak rate ratio for lung cancer among men was in the period 1992–1994, with an RR estimate of 3.61 (95% CI = 3.13 to 4.16), and the peak rate ratio for lung cancer among women was in the period 1989–1991, with an RR of 3.26 (95% CI = 2.50 to 4.23). We did a similar analysis for bladder cancer (Table 3). The peak rate ratio for bladder cancer among men was in the period 1986 − 1988, with an RR of 6.10 (95CI = 3.97 to 9.39), and the peak rate ratio among women was in the period 1992 − 1994, with an RR of 13.8 (95% CI = 7.74 to 24.5).

Mortality rates per 100.000 persons per year are also given in Table 2 for lung cancer and in Table 3 for bladder cancer. Combined lung and bladder cancer mortality rates in region II were highest in the period 1992–1994, with mortality rates per 100.000 persons of 153 for men (lung cancer, n = 130, plus bladder cancer, $n = 23$) in region II compared with 54 (lung cancer, $n = 47$, plus bladder cancer, $n = 7$) in region V. The corresponding rates for women in 1992 − 1994 were 50 in region II (lung cancer, $n = 34$, plus bladder cancer, n = 16) compared with 19 in region V (lung cancer, $n = 17$, plus bladder cancer, $n = 2$).

The time patterns of increased mortality from lung and bladder cancers in region II compared with region V are displayed in Fig. 2. Separately estimated mortality rate ratios for each 3-year period fall mostly within the confidence bands of the smoothed Poisson regression functions. Trends of increasing risk are generally apparent between about 1968 (10 years after high exposures com- menced) and 1978 (20 years after high exposures commenced). The rate of lung cancer in men in region II was already about twice that in region V by the period 1968 – 1970 (RR = 1.98, 95% CI = 1.57 to 2.51; Table 2). Among women, the lung cancer RR had reached 2.89 (95% CI = 2.00 to 4.18) by the period 1977–1979. Bladder cancer mortality rate ratios rose even higher than those for lung cancer (Fig. 2). Among men, the RR had reached 5.95 (95% CI = 2.22 to 16.0) in the period 1974–1975, and among women, the RR had reached 3.45 (95% CI = 1.34 to 8.91) in the period 1971–1973 (Table 3).

With the exception of bladder cancer among women, there was evidence that rate ratios had started to decline after peaking around 1990 (Fig. 2). However, the wide confidence bands preclude definitive statements about reductions in rate ratios, especially among women.

The time trends in the rate ratios for three birth cohorts comparing region II with region V are presented in Fig. 3 . Lung cancer rate ratios were markedly elevated for the male birth cohort born in 1938 – 1957, who would have experienced high exposures as young children. By contrast, there was no evident difference in lung cancer rate ratios between the birth cohorts of women. For bladder cancer, high rate ratios can be seen for each birth cohort. The mortality rate ratio appeared to continue to increase for women born before 1918, but by the year 2000, the data involved a relatively small number of women, all above the age of 80 years.

3 Discussion

In this study, clear latency patterns between arsenic exposure and lung and bladder cancer mortality can be seen because of the large population exposed (251 976 residents in region II in 1970), accurate data on past exposure, and the precise time pattern of commencement and decline of high exposures. To highlight the size of the study, there were 3406 lung cancer deaths in the exposed population (see Table 2). Latency patterns are usually difficult to obtain for human cancers. The largest study so far published concerning lung cancer and arsenic involved 1525 lung cancer deaths, but that number includes both those exposed to arsenic and those not exposed to arsenic (26). We report here very high rate ratios for both lung cancer (3- to 4-fold) and bladder cancer (6- to 10-fold) following exposure to arsenic in drinking water. Rate ratio estimates of this size for a defined large population living in a region of a country are, we believe, without precedent for any cause of any human cancer. Active cigarette smoking results in higher relative risks for lung cancer among smokers compared with nonsmokers (10- to 20-fold) (27) but lower relative risk estimates for bladder cancer (2- to 4-fold) (28). However, relative risks with active cigarette smoking relate to the subset of a population who are smokers rather than involving a total population within a region of a country, as is the case with arsenic in water sources in region II of Chile, so the local public health impact of arsenic in water is potentially greater.

The birth cohort analyses show very high lung cancer relative risks for men born in the 20-year period, 1938–1957, just before the very high arsenic exposures commenced, and who would have been exposed as children or adolescents (Fig. 3). We previously reported high lung cancer mortality rates for young adults in the period 1989–2000 in Antofagasta who would have experienced early life exposure to arsenic in water (29). The data presented here include all of region II, span more years, and confirm the earlier findings. We know of no mechanistic explanation for the finding that boys appear to be more susceptible than girls to subsequently developing lung cancer when exposed to arsenic in water as children. We are planning further investigations to assess variation in the impact of exposure at different ages.

Table 1. Arsenic concentrations (µ g/L) in drinking water in major cities and towns in region II of Chile and population-weighted averages for all of region II from 1950 to 1994 calculated using 1991 census population numbers

City or town					Period				
(population, 1991 census)	1950-54	1955-59	1960-6	1965-69	1970-74	1975-79	1980-84	1985-89	1990-94
Antofagasta (219 310)	90	870	870	870	260	110	80	60	40
Mejillones (6 134)	90	870	870	870	260	110	80	60	40
Calama (100 283)	120	120	120	120	240	230	110	80	40
Chuquicamata (17 414)	250	150	130	130	130	110	80	60	10
Tocopilla (21 039)	250	250	250	250	520	460	110	80	40
María Elena (15 470)	250	250	250	250	520	460	110	80	40
Taltal (7 620)	60	60	60	60	60	60	60	60	60
San Pedro (3 070)	600	600	600	600	600	600	600	600	600
Average for region II	123	569	568	568	272	176	94	71	43

Data published previously (15).

A major strength of our study was the ability to assess arsenic exposures with less uncertainty than other studies. In almost all epidemiology studies of arsenic-related cancers that have been carried out to date, a major problem has been retrospective exposure assessment. High arsenic concentrations in drinking water in the rest of the world (including Taiwan, Argentina, Mongolia, Bangladesh, India, Mexico, Thailand, Nepal, and the United States) are found in well water sources, for which there are, at best, limited historical records of arsenic concentrations (30).

Fig. 2. Lung and bladder cancer mortality rate ratios comparing region II with region V for men and women aged 30 and above, separately, as estimated by Poisson regression with smoothing. The shading represents the 95% confidence bands. The circles represent the mortality rate ratios plotted at the midpoint of each successive 3-year period. Histograms (gray lines) of the population-weighted average arsenic water concentrations for region II, from 1950 to 1994 in 5-year increments, are also presented (vertical axes at right).

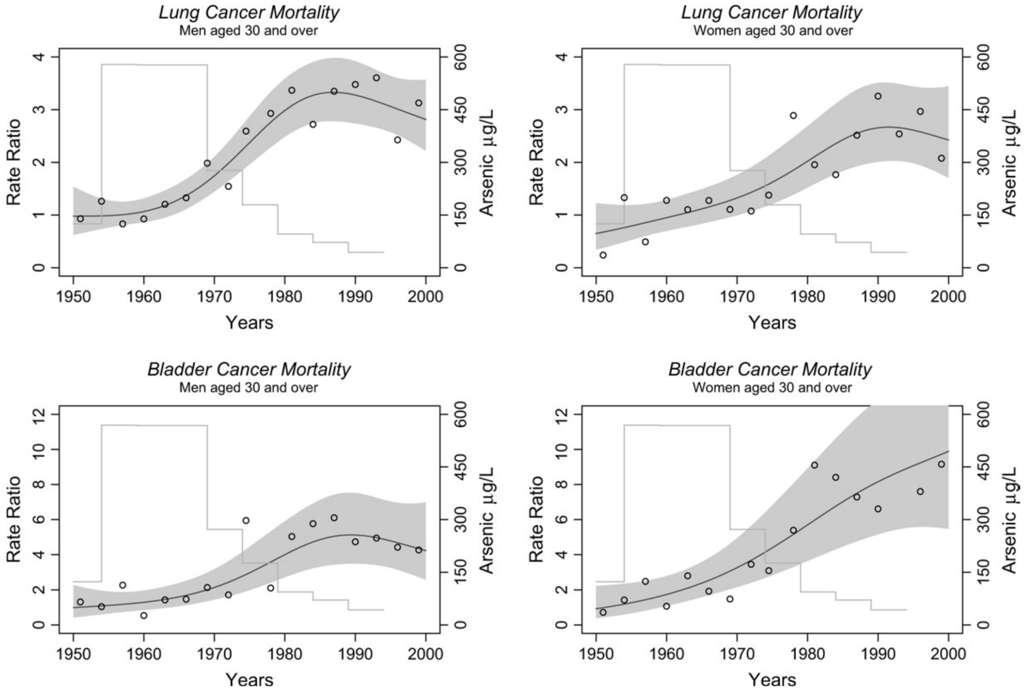


Table 2. Observed lung cancer deaths and lung cancer mortality rates and rate ratios for men and women aged 30 years and above in region II compared with region V and the rest of Chile, 1950–2000

Years	Males						Females					
	No. of deaths		Mortality rates (per 100.000)		Rate ratio (95% CI) compared with		No. of deaths		Mortality rates (per 100.000)		Rate ratio (95% CI) compared with	
	II	V	II	V	V	Rest of Chile[†]	II	V	II	V	V	Rest of Chile[†]
1950–1952	22	80	20	24	0.93 (0.59 to 1.46)		2	40	2	11	0.24 (0.0 to 14.64)	
1953–1955	38	103	33	29	1.26 (0.87 to 1.83)		12	43	13	11	1.33 (0.70 to 2.52)	
1956–1958	26	110	21	29	0.83 (0.54 to 1.27)		4	39	4	9	0.49 (0.18 to 1.37)	
1959–1961	35	137	28	33	0.93 (0.64 to 1.34)		13	49	12	11	1.28 (0.69 to 2.35)	
1962–1964	48	154	37	35	1.20 (0.87 to 1.66)		12	54	10	11	1.10 (0.59 to 2.06)	
1965–1967	45	139	34	30	1.33 (0.95 to 1.86)		16	64	13	12	1.28 (0.74 to 2.21)	
1968–1970	102	222	75	44	1.98 (1.57 to 2.51)		18	85	14	15	1.11 (0.67 to 1.84)	
1971–1973	109	312	75	58	1.54 (1.24 to 1.92)	2.46 (2.02 to 2.99)	23	112	16	18	1.08 (0.69 to 1.69)	1.79 (1.18 to 2.72)
1974–1975[†]	95	163	92	43	2.59 (2.01 to 3.34)	3.42 (2.77 to 4.22)	14	53	14	12	1.38 (0.76 to 2.48)	1.68 (0.99 to 2.87)
1977–1979	175	268	103	43	2.93 (2.42 to 3.55)	3.46 (2.96 to 4.03)	44	79	26	11	2.89 (2.00 to 4.18)	2.88 (2.12 to 3.91)
1980–1982	238	319	131	48	3.37 (2.85 to 3.98)	3.96 (3.47 to 4.53)	42	111	23	14	1.96 (1.37 to 2.79)	2.23 (1.64 to 3.05)
1983–1985	209	347	104	48	2.72 (2.29 to 3.23)	3.04 (2.64 to 3.51)	40	116	20	14	1.77 (1.23 to 2.53)	1.85 (1.35 to 2.54)
1986–1988	251	338	114	43	3.35 (2.84 to 3.94)	3.74 (3.28 to 4.26)	66	133	30	15	2.52 (1.87 to 3.38)	2.80 (2.18 to 3.59)
1989–1991	315	408	131	49	3.48 (3.00 to 4.03)	4.13 (3.68 to 4.64)	92	142	38	15	3.26 (2.50 to 4.23)	3.41 (2.76 to 4.22)
1992–1994	345	425	130	47	3.61 (3.13 to 4.16)	4.20 (3.76 to 4.70)	91	178	34	17	2.54 (1.97 to 3.27)	2.81 (2.27 to 3.48)
1995–1997	302	534	101	55	2.43 (2.11 to 2.79)	3.27 (2.91 to 3.68)	121	199	42	18	2.97 (2.37 to 3.72)	3.16 (2.62 to 3.80)
1998–2000	345	460	104	44	3.13 (2.72 to 3.60)	3.63 (3.25 to 4.05)	96	222	31	19	2.08 (1.64 to 2.64)	2.26 (1.84 to 2.77)

** CI = confidence interval; R II = region II; R V = region V.*

† Mortality data for the rest of Chile were not available in electronic form until 1971.

‡ Omitting 1976, for which data were not available in electronic form.

Table 3. Observed bladder cancer deaths and bladder cancer mortality rates and rate ratios for men and women aged 30 years and above in region II compared with region V and the rest of Chile for the period 1950–2000*

Years	Males						Females					
	No. of deaths		Mortality rates (per 100.000)		Rate ratio (95% CI) compared with		No. of deaths		Mortality rates (per 100.000)		Rate ratio (95% CI) compared with	
	II	V	II	V	V	Rest of Chile[†]	II	V	II	V	V	Rest of Chile[†]
1950–1952	5	14	4	4	1.31 (0.16 to 11.0)		1	7	1	2	0.71 (–)	–
1953–1955	4	14	3	4	1.04 (0.34 to 3.16)		2	7	2	2	1.42 (0.29 to 6.83)	
1956–1958	8	13	7	3	2.27 (0.94 to 5.47)		3	6	3	1	2.48 (0.62 to 9.91)	
1959–1961	3	21	2	5	0.53 (0.16 to 1.79)		3	14	3	3	1.07 (0.31 to 3.71)	
1962–1964	5	14	4	3	1.42 (0.51 to 3.96)		6	11	5	2	2.80 (1.03 to 7.57)	
1965–1967	9	26	7	6	1.47 (0.69 to 3.13)		4	11	3	2	1.92 (0.61 to 6.03)	
1968–1970	11	23	8	5	2.13 (1.04 to 4.37)		3	11	2	2	1.47 (0.41 to 5.28)	
1971–1973	9	24	6	4	1.71 (0.80 to 3.69)	3.64 (1.82 to 7.31)	7	11	5	2	3.45 (1.34 to 8.91)	6.80 (3.01 to 15.4)
1974–1975[‡]	9	7	9	2	5.95 (2.22 to 16.0)	5.00 (2.50 to 10.0)	4	7	4	1	3.09 (0.9 to 10.56)	5.35 (1.90 to 15.1)
1977–1979	17	38	10	6	2.10 (1.19 to 3.72)	2.91 (1.78 to 4.76)	10	10	6	2	5.39 (2.24 to 13.0)	4.71 (2.45 to 9.05)
1980–1982	35	33	19	5	5.04 (3.13 to 8.10)	5.10 (3.59 to 7.26)	22	13	12	2	9.10 (4.59 to 18.1)	6.28 (4.01 to 9.82)
1983–1985	41	34	20	5	5.77 (3.66 to 9.09)	5.42 (3.91 to 7.51)	22	14	11	2	8.41 (4.30 to 16.4)	7.14 (4.54 to 11.2)
1986–1988	47	37	21	5	6.10 (3.97 to 9.39)	5.99 (4.41 to 8.14)	37	27	17	3	7.28 (4.44 to 12.0)	10.4 (7.23 to 14.9)
1989–1991	52	53	22	6	4.73 (3.23 to 6.94)	5.86 (4.38 to 7.84)	35	28	14	2	6.61 (4.02 to 10.9)	8.33 (5.80 to 12.0)
1992–1994	62	60	23	7	4.95 (3.47 to 7.06)	6.63 (5.06 to 8.67)	42	16	17	2	13.8 (7.74 to 24.5)	9.32 (6.67 to 13.0)
1995–1997	56	59	19	6	4.13 (3.07 to 6.38)	5.05 (3.82 to 6.67)	44	30	15	2	7.60 (4.78 to 12.1)	8.45 (6.11 to 11.7)
1998–2000	58	62	18	6	4.27 (2.98 to 6.11)	5.12 (3.89 to 6.73)	50	28	16	2	9.16 (5.76 to 14.5)	8.53 (6.29 to 11.6)

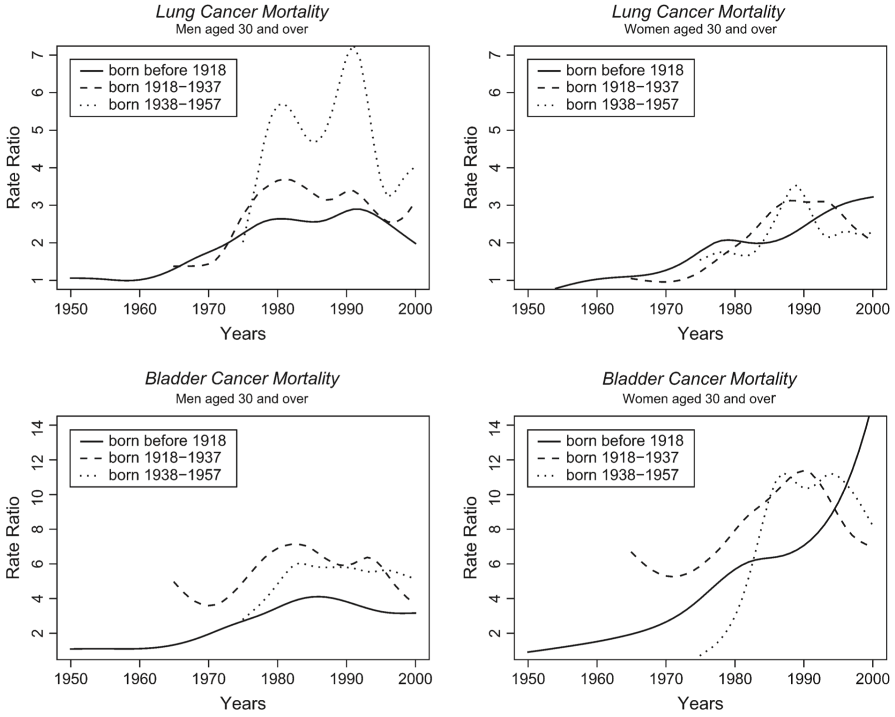
** CI = confidence interval; R II = region II; R V = region V.*

† Mortality data for the rest of Chile were not available in electronic form until 1971.

‡ Omitting 1976, for which data were not available in electronic form.

Many of the water sources in these populations are small private domestic wells used by only a few people or families. Thus, assessing exposure in epidemiology studies can involve identifying and measuring arsenic concentrations in water from hundreds of individual wells. There may be considerable variation in arsenic concentration among wells close to each other, and there is often uncertainty about the well from which a person consumed water decades ago. This lack of information leads to major uncertainties in individual and population exposure estimates for long latency outcomes, such as cancer. We have carried out studies involving sampling of well water in Argentina (12), West Bengal (31; 32; 33; 34; 35), and California and Nevada (11; 36; 37), and in each location there have been major problems in locating (for sampling) the wells used by individuals decades ago. Even if the right well is located, there is uncertainty about whether current arsenic levels are representative of the water consumed at the earlier period. In the epidemiology studies from Taiwan, for example, past exposure to arsenic was not determined, other than to note how long individuals drank from the well they were using at the time the studies were con- ducted. In one recent Taiwan study, for example, only a single measurement was used for each subject, even if subjects had used many different water sources over the course of their lives (14).

Fig. 3. Lung and bladder cancer mortality rate ratios comparing region II with region V for men and women aged 30 and above, separately, as estimated by Poisson regression with smoothing for three birth cohorts, one born in 1938 – 1957 before the high exposures commenced in 1958 (dotted lines) and two older birth cohorts, one born before 1918 (solid lines) and the second born in 1918–1937 (dashed lines).



Estimating exposures to arsenic in drinking water in Northern Chile is considerably less uncertain than doing so in the places described above. In Northern Chile, wells were not used to obtain water. Until recently, bottled water was also not used. Because of the extreme dryness of the area, all water came from a relatively small number of large municipal water supplies, for which there are historical records of water concentrations of arsenic. As a result, simply knowing the town in which a person lived during a particular year accurately establishes the arsenic concentration in the water they drank. However, in this paper, we present mortality findings for the region as a whole, rather than subdividing by city and town, because census and mortality data were not available by city and town of residence for the whole period 1950–2000.

We have previously shown that the elevated cancer rates found in region II are related to arsenic contamination of water supplies and not to some bias or confounding factor, such as smoking (15). While smoking is an established cause of cancers of the lung and bladder, confounding due to smoking can be dismissed as the reason for the increased mortality from these cancers in region II, for three reasons. First, as noted above, the smoking data from region II do not support higher smoking rates than in the rest of Chile. Second, the extent of increased risks is much too large to attribute to cigarette smoking. Studies in various populations have shown that the relative risks of bladder cancer for smokers compared with nonsmokers are generally in the range of 2–4 (28). On this basis alone, smoking can be dismissed as the reason for the bladder cancer mortality ratios shown in Table 3, many of which exceed 5. Smokers do have increased mortality from lung cancer with relative risks of the order of 10–20 when compared with nonsmokers (27), and it might seem that the standardized mortality ratios reaching around 3–4 for men and women in region II could be due to smoking. However, this is not the case because smoking also occurs in the comparison populations, region V and the rest of Chile. In fact, it is extremely unlikely that confounding due to smoking could result in lung cancer rate ratios greater than 2 (38). Third, the case–control study of lung cancers diagnosed in 1994–1996 in region II of Chile had data on individual smoking for each participant, and although there was strong evidence of increased risks with arsenic exposure, there was no evidence of confounding with smoking (16).

The latency patterns we show here provide further evidence for the causal relationship between arsenic in the water in region II of Chile and increased rates of lung and bladder cancer because increased rates of these cancers temporally followed the increase in arsenic exposure in a plausible manner. Smoking is the most important population cause for both lung cancer and bladder cancer in most of the world. As noted above, the smoking prevalence figures given above show that smoking rates in region II and region V are about the same.

Although confounding from different smoking patterns is not an issue, potential synergy between arsenic and smoking could be important. Other research we have conducted (9; 10; 11) suggests that smoking might be a cofactor with arsenic in bladder cancer causation and that smoking and arsenic might be synergistic in increasing the risk of lung cancer (16). Historically, in Chile as elsewhere, males have been more likely to be smokers than females. It is possible that interaction of arsenic with tobacco results in some of the differences in the mortality trends seen in our study between males and females. However, the mortality rate ratio for bladder cancer rises higher for women than men, which does not support the idea that smoking is a required co-carcinogen for arsenic to cause bladder cancer. If smoking were a required co-carcinogen, then one would expect the impact of arsenic on bladder cancer rates to be greater for men than for women since they smoke more.

The increase in bladder cancer mortality we report in this study could be due in part to increased fatality of tumors related to arse- nic and not just to increased incidence alone. Lung cancer is highly fatal, and trends in mortality rates are closely related to incidence rates. However, bladder cancer survival rates are relatively good, and bladder cancer mortality rates therefore reflect a combination of both incidence rates and case fatality rates. Our previous inves- tigation of chromosomal alterations in bladder cancer biopsies, including tumor biopsies from region II of Chile, indicated that tumors from arsenic-exposed patients may behave more aggres- sively than tumors from unexposed patients (39).

A limitation of this ecologic study is that it could not account for migration in and out of region II. However, because arsenic exposures in region II are much higher than in the rest of Chile and elsewhere, migration in or out would have diluted, not increased, the rate ratios reported in this paper. Patterns of increased mortality are clear, and rate ratios would likely have increased even further if analyses could have been confined to persons with long-term residence in region II. In addition, migration among regions in Chile is relatively low. From 1965 to 2000, annual internal migration between regions was only 0.6%, compared with 1.2% in Argentina, 3.1% in the United Kingdom, and 6.6% in the United States (40). Further limitations include not having individual data on arsenic exposure and not having individual data on other risk factors such as smoking and occupational exposures.

In conclusion, we have found a clear latency pattern for lung and bladder cancer mortality for both men and women that is consistent with the effects of a large increase in population exposure to arsenic starting in 1958. Increased rate ratios became evident close to 10 years after exposure increased, peaked in the years around 1990, and continued to be markedly elevated up to the year 2000. The impact of arsenic in drinking water on this large population is without precedent for environmental causes of human cancer, and it points to the public health priority of ensuring that arsenic concentrations in drinking water are controlled worldwide.

Bibliography

[1] Cornejo-Catalan J. Panorama de arsenicismo en la II region. In: Primera Jornada Sobre Arsenicismo Laboral y Ambiental II Region. Antofagasta (Republica de Chile): Ministerio de Salud Antofagasta Departamento de Programas Sobre El Ambiente, y Asociacion Chilena se Seguridad; 1991. p. 15–34.

[2] Chen CJ, Chuang YC, Lin TM, Wu HY. Malignant neoplasms among residents of a blackfoot disease-endemic area in Taiwan: high-arsenic artesian well water and cancers. Cancer Res 1985; 45: 5895–9.

[3] Chen CJ, Wang CJ. Ecological correlation between arsenic level in well water and age-adjusted mortality from malignant neoplasms. Cancer Res 1990; 50: 5470–4 .

[4] Chen CJ, Wu MM, Lee SS, Wang JD, Cheng SH, Wu HY. Atherogenicity and carcinogenicity of high-arsenic artesian well water. Multiple risk factors and related malignant neoplasms of blackfoot disease. Arteriosclerosis 1988; 8: 452–60 .

[5] Chiou HY, Hsueh YM, Liaw KF, Horng SF, Chiang MH, Pu YS, et al. Incidence of internal cancers and ingested inorganic arsenic: a seven-year follow-up study in Taiwan. Cancer Res 1995; 55: 1296–300 .

[6] Hopenhayn-Rich C, Biggs ML, Fuchs A, Bergoglio RM, Tello EE, Nicolli HB, et al. Bladder cancer mortality associated with arsenic in drinking water in Argentina. Epidemiology 1996; 7: 117–24.

[7] Hopenhayn-Rich C, Biggs ML, Smith AH. Lung and kidney cancer mortality associated with arsenic in drinking water in Cordoba, Argentina. Int J Epidemiol 1998; 27: 561–9 .

[8] International Agency for Research on Cancer. Some drinking-water disinfectants and contaminants, including arsenic. Vol 84. Lyon (France): World Health Organization; 2004.

[9] National Research Council. Arsenic in drinking water: 2001 update. Washington (DC): National Academy Press; 2001.

[10] Bates MN, Smith AH, Cantor KP. Case-control study of bladder cancer and arsenic in drinking water. Am J Epidemiol 1995; 141: 523–30 .

[11] Steinmaus C, Yuan Y, Bates MN, Smith AH. Case-control study of bladder cancer and drinking water arsenic in the Western United States. Am J Epidemiol 2003; 158: 1193–201.

[12] Bates MN, Rey OA, Biggs ML, Hopenhayn C, Moore LE, Kalman DA, et al. Case-control study of bladder cancer and exposure to arsenic in Argentina. Am J Epidemiol 2004; 15: 381–9 .

[13] Chen CJ, Chuang YC, You SL, Lin TM, Wu HY. A retrospective studyon malignant neoplasms of bladder, lung and liver in blackfoot disease endemic area in Taiwan. Br J Cancer 1986; 53: 399–405.

[14] hiou HY, Chiou ST, Hsu YH, Chou YL, Tseng CH, Wei ML, et al. Incidence of transitional cell carcinoma and arsenic in drinking water: a follow-up study of 8,102 residents in an arseniasis-endemic area in north- eastern Taiwan. Am J Epidemiol 2001; 153: 411–8 .

[15] Smith AH , Goycolea M, Haque R, Biggs ML. Marked increase in bladder and lung cancer mortality in a region of Northern Chile due to arsenic in drinking water. Am J Epidemiol 1998; 147: 660–9.

[16] Ferreccio C, Gonzalez CA, Milosavjlevic V, Marshall G, Sancha AM, Smith AH. Lung cancer and arsenic concentrations in drinking water in Chile. Epidemiology 2000; 11: 673–9 .

[17] Instituto Nacional de Estadística. Anuario de Demografia 1995. Santiago (Chile): Departamento de Estadisticas, Demograficas y Sociales, Servicio de Registro Civil e Identificacion. Ministerio de Salud; 1996.

[18] La Encuesta de Caracterización Socioeconómica. Ministerio de Planifi cacion y Cooperacion Nacional Republica de Chile. Illa Encuestra CASEN (Caracterizacion Socio Economica Nacional). 15th ed. Santiago (Chile); 1992.

[19] Castillo B, Mardones G. Medical certifi cation of deaths in the health services of Chile . Rev Med Chil 1986; 114: 693–700 .

[20] World Health Organization . Manual of the international statistical classification of diseases, injuries and causes of death, ninth revision. Vol 1. Geneva (Switzerland): World Health Organization; 1977.

[21] Lachin JM. Biostatistical methods: the assessment of relative risks. New York: John Wiley; 2000. p. 476–8 .

[22] Hastie TJ , Tibshirani T. Generalized additive models. London: Chapma and Hall; 1990.

[23] Roemer WH , van Wijnen JH . Daily mortality and air pollution along busy streets in Amsterdam, 1987–1998 . Epidemiology 2001; 12: 649–53.

[24] Schwartz J. Air pollution and daily mortality in Birmingham, Alabama . Am J Epidemiol 1993; 137: 1136–47.

[25] Hopenhayn C, Ferreccio C, Browning SR, Huang B, Peralta C, Gibb H, et al . Arsenic exposure from drinking water and birth weight. Epidemiology 2003; 14: 593–602 .

[26] Chiu HF, Ho SC, Yang CY. Lung cancer mortality reduction after installation of tap-water supply system in an arseniasis-endemic area in South-western Taiwan. Lung Cancer 2004; 46: 265–70 .

[27] Halpern MT, Gillespie BW, Warner KE. Patterns of absolute risk of lung cancer mortality in former smokers. J Natl Cancer Inst 1993; 85: 457–64.

[28] Silverman DT, Hartge P, Morrison AS, Devesa SS. Epidemiology of bladder cancer . Hematol Oncol Clin North Am 1992; 6: 1–30.

[29] Smith AH, Marshall G, Yuan Y, Ferreccio C, Liaw J, von Ehrenstein O, et al. Increased mortality from lung cancer and bronchiectasis in young adults after exposure to arsenic in utero and in early childhood. Environ Health Perspect 2006; 114: 1293–6.

[30] Nordstrom DK. Public health. Worldwide occurrences of arsenic in ground water. Science 2002; 296: 2143–5.

[31] Mitra SR, Mazumder DN, Basu AR, Block GS, Haque R, Samanta S, et al. Nutritional factors and susceptibility to arsenic-caused skin lesions in West Bengal, India . Environ Health Perspect 2004; 112: 1104–9 .

[32] von Ehrenstein OS, Guha Mazumder DN, Yuan Y, Samanta S, Balmes J, Sil A, et al. Decrements in lung function related to arsenic in drinking water in West Bengal, India. Am J Epidemiol 2005; 162: 533–41.

[33] Guha Mazumder DN, Steinmaus C, Bhattacharya P, von Ehrenstein OS, Ghosh N, Gotway M, et al. Bronchiectasis in persons with skin lesions resulting from arsenic in drinking water. Epidemiology 2005; 16: 760–5.

[34] von Ehrenstein OS, Guha Mazumder DN, Hira-Smith M, Ghosh N, Yuan Y, Windham G, et al. Pregnancy outcomes, infant mortality and arsenic in drinking water in West Bengal, India. Am J Epidemiol 2006; 163: 662–9.

[35] Haque R, Mazumder DN, Samanta S, Ghosh N, Kalman DA, Smith MM, et al. Arsenic in drinking water and skin lesions: dose-response data from West Bengal, India. Epidemiology 2003; 14: 174–82.

[36] Warner ML, Moore LE, Smith MT, Kalman DA, Fanning E, Smith AH. Increased micronuclei in exfoliated bladder cells of individuals who chronically ingest arsenic-contaminated water in Nevada. Cancer Epidemiol Biomarkers Prev 1994; 3: 583–90.

[37] Steinmaus CM, Yuan Y, Smith AH. The temporal stability of arsenic concentrations in well water in western Nevada. Environ Res 2005; 99: 164–8.

[38] Axelson O. Aspects on confounding in occupational health epidemiology. Scand J Work Environ Health 1978; 4: 85–9.

[39] Moore LE, Smith AH, Eng C, Kalman DA, DeVries S, Bhargava V, et al. Arsenic-related chromosomal alterations in bladder cancer. J Natl Cancer Inst 2002; 94: 1688–96.

[40] Soto R, Torche A. Spatial inequality, migration, and economic growth in Chile. Cuad Econ 2004; 41: 401–24.

Article 3.2

Increased Mortality from Lung Cancer and Bronchiectasis in Young Adults after Exposure to Arsenic in Utero and in Early Childhood

Allan H. Smith, Guillermo Marshall, Yan Yuan, Catterina Ferreccio, Jane Liaw, Ondine von Ehrenstein, Craig Steinmaus, Michael N. Bates, and Steve Selvin

School of Public Health, University of California, Berkeley, CA and Universidad Católica de Chile

Abstract. Arsenic in drinking water is an established cause of lung cancer, and preliminary evidence suggests that ingested arsenic may also cause nonmalignant lung disease. Antofagasta is the second largest city in Chile and had a distinct period of very high arsenic exposure that began in 1958 and lasted until 1971, when an arsenic removal plant was installed. This unique exposure scenario provides a rare opportunity to investigate the long-term mortality impact of early-life arsenic exposure. In this study, we compared mortality rates in Antofagasta in the period 1989–2000 with those of the rest of Chile, focusing on subjects who were born during or just before the peak exposure period and who were 30–49 years of age at the time of death. For the birth cohort born just before the high-exposure period (1950–1957) and exposed in early childhood, the standardized mortality ratio (SMR) for lung cancer was 7.0 [95dence interval (CI), 5.4–8.9; p ¡ 0.001] and the SMR for bronchiectasis was 12.4 (95p ¡ 0.001). For those born during the high-exposure period (1958–1970) with probable exposure in utero and early childhood, the corresponding SMRs were 6.1 (95cancer and 46.2 (95greatly increasing subsequent mortality in young adults from both malignant and nonmalignant lung disease. Key words: arsenic, bronchiectasis, childhood exposure, chronic obstructive pulmonary disease, drinking water, in utero exposure.[1]

The International Agency for Research on Cancer (IARC) has classified arsenic in drinking water as a group 1 carcinogen that causes skin cancer, bladder cancer, and lung cancer (IARC 2002). Substantial evidence supports the biologic plausibility that exposure to arsenic can lead to skin and bladder cancer. For example, arsenic concentrates in the skin and is known to cause nonmalignant skin lesions [National Research Council (NRC) 2001], and the major pathway of excretion is in urine, giving plausibility to increased bladder cancer rates (NRC 2001). Although it is known that inhalation of arsenic may cause lung cancer, the findings of increased lung cancer mortality after ingestion in drinking water were unexpected because all other known lung carcinogens act via inhalation. However, the evidence based on multiple studies in Taiwan (Chen and Wang 1990; Chen et al. 1985, 1988; Wu et al. 1989), Chile (Ferreccio et al. 2000; Smith et al. 1998), Argentina (Hopenhayn-Rich et al. 1998), and Japan (Tsuda et al. 1989, 1995) is sufficient to conclude that there is a causal relationship. In fact, lung cancer is the main long-term cause of death from ingesting arsenic in drinking water (NRC 2001). In region II of Chile, which includes the city of Antofagasta, overall lung cancer mortality rates for men and women were previously

[1] Smith AH, Marshall G, Yuan Y, Ferreccio C, Liaw J, von Ehrenstein O, et al. 2006. Increased mortality from lung cancer and bronchiectasis in young adults after exposure to arsenic in utero and in early childhood. *Environ Health Perspect* 114:1293-1296.

found to be at least 3-fold higher than for the rest of Chile (Smith et al. 1998), and lung cancer relative risk estimates increased nearly 9-fold in those with the highest exposures (Ferreccio et al. 2000).

Several known lung carcinogens cause chronic nonmalignant respiratory diseases, including cigarette smoking, which causes chronic obstructive pulmonary disease (COPD); asbestos, which causes asbestosis; and silica, which causes silicosis. To date, however, relatively little attention has been given to whether or not ingestion of arsenic in drinking water causes nonmalignant pulmonary disease. The first reports of chronic respiratory symptoms came from small investigations in Antofagasta in the 1970s (Zaldivar 1974, 1977, 1980; Zaldivar and Ghai 1980). Before 1958, the water supply in the main city of Antofagasta had an arsenic concentration of about 90 µg/L. A growing population led to supplementation of Antofagasta's water supply in the late 1950s with water from rivers with arsenic concentrations near 1,000 µg/L. Because this area is among the driest places on Earth, there are very few individual water supplies, and almost everyone drinks water from the same municipal sources. After the installation of a new treatment plant in 1971, arsenic levels in Antofagasta water dropped abruptly to about 90 µg/L and have been progressively reduced further in recent years (Ferreccio et al. 2000). These data are shown in Figure 1.

In a 1998 publication concerning region II, increased COPD mortality was reported for the 30- to 39-year age group (Smith et al. 1998). Based on the time period in which mortality was assessed (1989–1993), subjects in the 30- to 39-year age group would have been in utero or young children at the time of the peak exposure period in Antofagasta. These results were based on a small number of cases but were later supported by findings from other arsenic-exposed regions. For example, increases in symptoms of chronic respiratory disease were found to be associated with arsenic ingestion in studies in West Bengal, India (De et al. 2004; Guha Mazumder et al. 2000) and Bangladesh (Milton and Rahman 2002). Recently, two studies in West Bengal involving participants with arsenic-caused skin lesions reported major deficits in pulmonary function (von Ehrenstein et al. 2005) and a 10-fold increase in prevalence of bronchiectasis identified by high-resolution computed tomography (Guha Mazumder et al. 2005).

The distinct period of high arsenic exposure in Antofagasta from 1958 through 1970 offers the opportunity to investigate the health effects of early-life arsenic exposure. In this study, we take advantage of this unique situation in order to assess adult mortality in those born during the high-exposure period who would have experienced exposure in utero as well as early childhood, and those born just before 1958 who would have experienced high exposure during childhood but not in utero.

1 Materials and Methods

We obtained computerized mortality data for 1989–2000 from the Ministry of Health for all 13 regions of Chile. Deaths were divided into two groups: those who were residents of Antofagasta and neighboring Mejillones, cities that have the same water source; and those who were residents in all regions of Chile other than region II, in which Antofagasta and Mejillones are located. Two birth cohorts were defined for this investigation: those born in the period 1958–1970 (probable in utero exposure if resident in Antofagasta/Mejillones) and those born in 1950–1957 (probable childhood exposure if born in Antofagasta/ Mejillones). Causes of death were coded according to the International Classification of Diseases, 9th Revision (ICD-9; World Health Organization 1978), including lung cancer (ICD-9 code 162) and chronic respiratory disease (ICD-9 codes 490, 491, 492, 494, and 496). We obtained annual estimates of the population living in Antofagasta/Mejillones in region II, and for the rest of Chile excluding region II, for 1989–2000 from the National Institute of Statistics (Instituto Nacional de Estadísticas) stratified by age and sex.

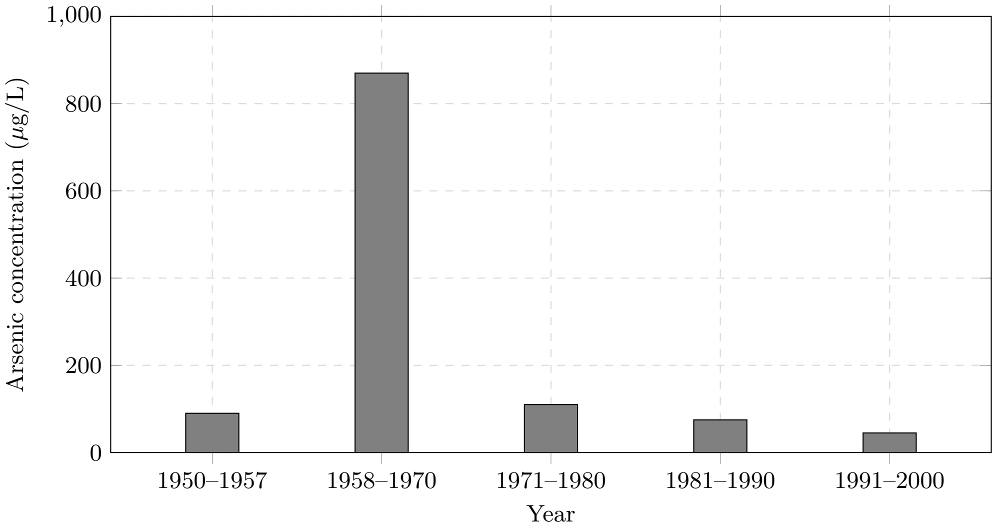


Fig. 1. Arsenic concentrations in Antofagasta/ Mejillones water by year. An arsenic removal plant was installed in 1971.

In 2000, the most recent year for which mortality data are available, the oldest persons in the first birth cohort born in the period 1950–1957 would have been 50 years old. We therefore calculated standardized mortality ratios (SMRs) for men and women separately, 30–49 years of age, using 10-year age groups (30–39 and 40–49 years) for standardization. Mortality in younger ages was not included because death from lung cancer or chronic respiratory disease is extremely rare in individuals ¡ 30 years of age. We calculated SMRs as the observed number of deaths divided by the expected number of deaths, using all regions in Chile outside of region II as the referent population. We estimated SMRs for lung cancer, for bronchiectasis, and for other COPD causes of death excluding bronchiectasis, and also for all other causes of death excluding lung cancer and COPD. We calculated tests of significance and confidence intervals (95% CIs) based on the Poisson distribution (Selvin 1995). In view of the clear direction of the a priori hypotheses for arsenic and both malignant and nonmalignant pulmonary diseases, we conducted one-tailed tests of significance for increases in these outcomes. We assessed tests for effect modification by age group (comparing 30–39 and 40–49 year age groups) and tests for effect modification by sex and for differences between those born in 1950–1957 and 1958–1970 by testing the pertinent Poisson regression interaction terms with two-tailed tests.

2 Results

SMRs for lung cancer and COPD are given in Table 1 for the 30–39 and 40–49 age groups separately and combined and for men and women separately and combined. Based on the Poisson regression interaction terms, there was no evidence of differences in rate ratios between 30–39 and 40–49 age groups for lung cancer and COPD causes of death, so we focused on the SMRs for the overall age range 30–49 years. For lung cancer, the SMR for 30–49 years of age was increased for those born in the period 1950–1957 for both men (SMR = 8.2; 95% CI, 6.2–10.8, $p < 0.001$) and women (SMR = 4.7; 95% CI, 2.7–7.7; $p < 0.001$). The lung cancer SMR was also increased for those born in 1958–1970 (women: SMR = 2.9; 95% CI, 0.6–8.5; $p = 0.087$; men: SMR = 8.1; 95% CI, 4.3–13.9; $p < 0.001$). Concerning COPD mortality, bronchiectasis SMRs were markedly increased for both men and women, especially for those born in the high-exposure period 1958–1970 (women: SMR = 50.1; 95% CI, 20.0–103; $p < 0.001$; men: SMR = 36.4; 95% CI, 4.1–132; $p = 0.001$). SMRs for other COPD causes of death excluding bronchiectasis were elevated, but much less than for bronchiectasis. Finally, for all other causes of death combined, there was little evidence of increased mortality for either birth cohort, as shown in Table 1.

The lung cancer relative risks are higher for men than for women, but the CIs for women are wide because of the relatively small numbers and overlap the lung cancer SMR for men (point estimate for men 30–49 years of age, 8.1; 95% CI for women, 0.6–8.5; Table 1). Testing Poisson regression interaction terms, there was little evidence of effect modification by sex for the period 1950–1957 ($p = 0.23$), but testing for effect modification for those born in the period 1958–1970 yields a p-value of 0.04, with higher relative risks for men than for women (8.1 for men and 2.9 for women). The pooled results are presented in Table 1 and Figure 2. They show that lung cancer rates are greatly increased for both those born in 1950–1957 with childhood exposure and for those born in 1958–1970 who would have experienced in utero exposure. However, for bronchiectasis, and to a lesser extent for other COPD mortality, the SMRs are much higher for those born in 1958–1970 (SMR = 46.2; 95% CI, 21.1–87.7; $p < 0.001$) than for those born in 1950–1957 before the very high exposures started (SMR = 12.4; 95% CI, 3.3–31.7; $p < 0.001$; Poisson regression test for difference in bronchiectasis rate ratios for the two periods, $p = 0.02$).

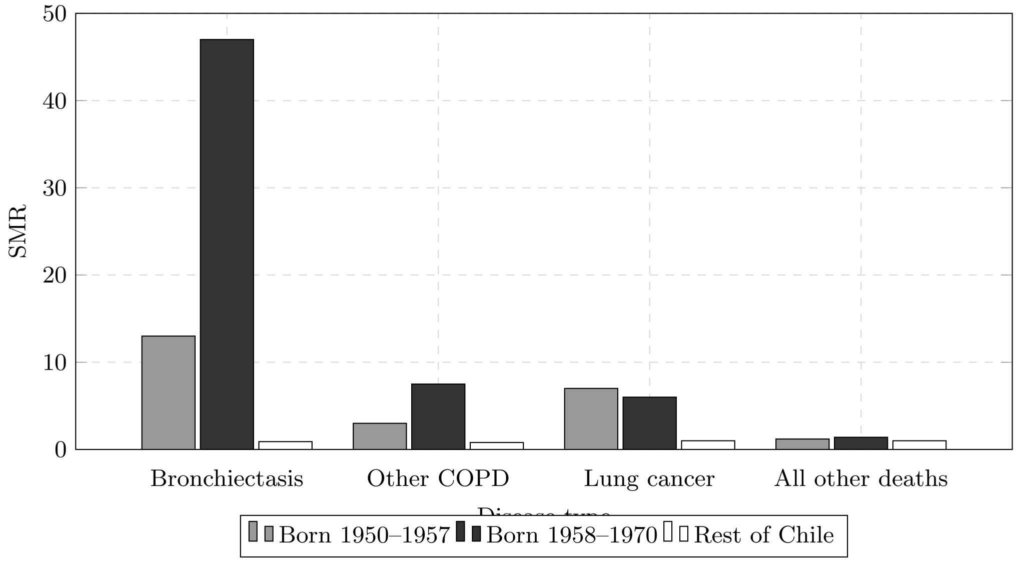


Fig. 2. COPD SMRs for Antofagasta/Mejillones for individuals 30–49 years of age, pooled.

Table 1. SMRs for bronchiectasis, other COPD, all other deaths, and lung cancer for Antofagasta/Mejillones, for ages 30–49, for men and women both separately and pooled.

Age (years)	Sex	Cause of death	Born 1950–1957				Born 1958–1970			
			O	E	SMR (95% CI)	*p*-Value	O	E	SMR (95% CI)	*p*-Value
30–39	Male	Lung cancer	15	1.17	12.8 (7.1–21.1)	<0.001	12	1.30	9.2 (4.8–16.1)	<0.001
		Bronchiectasis	3	0.15	19.4 (4.0–56.8)	0.001	2	0.05	36.4 (4.1–131.2)	0.001
		Other COPD	1	0.21	4.7 (0.1–26.0)	0.193	1	0.46	2.2 (0.1–12.1)	0.368
		All other deaths	129	155.78	0.8 (0.7–1.0)	0.987	305	304.41	1.0 (0.9–1.1)	0.494
	Female	Lung cancer	2	0.48	4.2 (0.5–15.1)	0.084	3	0.83	3.6 (0.7–10.5)	0.052
		Bronchiectasis	0	0.04	0.0	—	2	0.14	42.9 (15.7–93.4)	<0.001
		Other COPD	2	0.14	13.9 (1.7–50.2)	0.009	4	0.33	12.2 (3.3–31.2)	<0.001
		All other deaths	74	64.95	1.1 (0.9–1.4)	0.145	145	113.73	1.3 (1.1–1.5)	0.003
	Pooled	Lung cancer	17	1.65	10.3 (6.0–16.1)	<0.001	15	2.14	7.0 (3.9–11.6)	<0.001
		Bronchiectasis	3	0.19	15.8 (3.2–46.0)	0.001	4	0.19	41.1 (11.7–80.9)	<0.001
		Other COPD	3	0.36	8.4 (1.7–24.5)	0.006	5	0.79	6.3 (2.1–14.8)	0.001
		All other deaths	203	220.73	0.9 (0.8–1.1)	0.891	450	418.14	1.1 (1.0–1.2)	0.064
40–49	Male	Lung cancer	37	5.14	7.2 (5.1–9.9)	<0.001	1	0.29	3.4 (0.1–18.9)	0.255
		Bronchiectasis	0	0.10	0.0	—	0	0.0	0.0	—
		Other COPD	3	1.30	2.3 (0.5–6.7)	0.144	1	0.10	10.2 (0.3–56.8)	0.093
		All other deaths	270	292.37	0.9 (0.8–1.0)	0.911	21	19.66	1.1 (0.7–1.6)	0.411
	Female	Lung cancer	14	2.90	4.8 (2.6–8.1)	<0.001	0	0.20	0.0	—
		Bronchiectasis	1	0.04	27.6 (0.7–154)	0.036	0	0.0	0.0	—
		Other COPD	2	0.76	2.6 (0.3–9.5)	0.177	1	0.04	27.4 (0.7–153)	0.036
		All other deaths	178	147.78	1.2 (1.0–1.4)	0.009	17	11.92	1.4 (0.8–2.3)	0.097
	Pooled	Lung cancer	51	8.04	6.3 (4.7–8.3)	<0.001	1	0.50	2.0 (0.01–11.2)	0.391
		Bronchiectasis	1	0.13	7.5 (0.2–42.0)	0.124	1	0.0	0.0	—
		Other COPD	5	2.06	2.4 (0.8–5.7)	0.059	2	0.13	14.9 (1.8–53.7)	0.008
		All other deaths	448	440.15	1.0 (0.9–1.1)	0.361	38	31.58	1.2 (0.9–1.7)	0.146

Abbreviations: E, expected; O, observed.

Table 2. Smoking habits among men and women in the two major cities in region II in 1990 compared with data for the rest of Chile [no. (%)]

	Smoking habits (cigarettes/day)						
	Nonsmokers	Occasional	1–9	10–19	≥20	Unknown	Total
Antofagasta	163,500 (76.4)	13,223 (6.2)	27,445 (12.8)	7,845 (3.7)	1,800 (0.8)	270 (0.1)	214,083 (100)
Calama	92,214 (80.4)	8,268 (7.2)	10,944 (9.5)	1,788 (1.6)	1,233 (1.1)	222 (0.2)	114,669 (100)
Rest of Chile	5,443,466 (75.1)	581,686 (8.0)	837,878 (11.6)	228,617 (3.2)	109,421 (1.5)	46,215 (0.6)	7,247,283 (100)

Data were obtained from the Ministerio de Planificación y Coordinación Nacional República de Chile IIIa, *Encuesta Caracterización Socio Económica Nacional* (1992).

3 Discussion

Region II of Chile provides a unique opportunity to investigate arsenic health effects. It is one of the driest areas of the world, and water used in major cities and towns comes from single sources with known arsenic concentration. Furthermore, there was an abrupt onset of high exposure in 1958 in Antofagasta, the major city of region II with a population at that time of about 200,000 (Zaldivar 1974), and an abrupt reduction in exposure in 1971 when the first large arsenic removal plant in the world was installed there. Such clear-cut exposure patterns are rare in environmental epidemiology, except perhaps radiation exposure from use of the atomic bomb in Hiroshima and Nagasaki and, to a lesser extent, ionizing radiation from accidents at nuclear reactors.

The magnitude of the effects found on lung cancer and bronchiectasis mortality has no parallel with effects of other environmental exposures occurring in utero and/or in early childhood. No lung cancer cases were reported in 40 years among the in utero–exposed sur- vivors of the atomic bombing of Hiroshima and Nagasaki (Yoshimoto et al. 1988). Children with the highest gamma radiation exposure in Hiroshima and Nagasaki ¡ 10 years of age did not experience increased lung cancer risks as adults, but those exposed in the age range of 10–19 years had lung cancer relative risk estimates of about 2.5 those of young adults 30–39 years of age (Shimizu et al. 1990, figure 2). The evidence for an effect of child- hood exposure to environmental tobacco smoke on adult lung cancer rates is mixed, with a meta-analysis finding no overall evidence of increased risks (Boffetta et al. 2000). However, a prospective study reported a relative risk estimate of 3.6 (95% CI, 1.2–11.1) based on four lung cancer cases among those with “many hours” of daily exposure (Vineis et al. 2005). By contrast, we report here a total of 84 deaths from lung cancer after childhood exposure to high concentrations of arsenic in drinking water in Chile, a 6- and 7-fold increase above rates in the rest of Chile (Table 1).

Some supportive evidence provides biologic plausibility for arsenic having effects in utero. Arsenic crosses the placenta in animals and humans, and there is human evidence that arsenic is a developmental toxicant affecting birth weight and reproductive out- comes (Concha et al. 1998; Hanlon and Ferm 1987; Hopenhayn et al. 2003; Hopenhayn- Rich et al. 1999, 2000). A study conducted in Bangladesh showed an increased risk for stillbirth [odds ratio (OR) = 2.5; 95% CI, 1.5–4.9] and spontaneous abortion (OR = 2.5; 95% CI, 1.5–4.3) in women with current arsenic exposure $\geq$ 100 µg/L in water (Milton et al. 2005), and a study in West Bengal found increased risks of stillbirths (OR = 6.1; 95% CI, 1.5–24.0) (von Ehrenstein et al. 2006). As a whole, these epidemiologic data provide evi- dence that arsenic exposure in utero could be associated with a number of adverse effects. The present study, however, is the first to pro- vide evidence that early-life exposures may produce effects manifesting in adults.

Oral-dose animal studies demonstrate arsenic teratogenicity (Chattopadhyay et al. 2002; Vahter 1994). Of particular relevance to our study is evidence that arsenic is a transplacental carcinogen in mice (Waalkes et al. 2000). Female offspring of pregnant mice that were given high doses of arsenic in their drinking water developed tumors at multiple sites, including the lung, with lung carcinoma increased to 5 of 24 (21%) compared with 0 of 25 (0%) in the unexposed controls.

Strengths of our study include the extensive documentation of arsenic in drinking water in the Antofagasta water system. Records of arsenic levels in Antofagasta have been kept for the last 50 years, and almost all residents drink from the same water supply. One potential limitation of this study is that it is ecologic in nature, because overall mortality rates in the cities of Antofagasta/Mejillones were compared with those of the rest of Chile. Residence was determined from death certificates and relates to residence at the time at death. We cannot be certain that those manifesting the increased mortality were actually born in Antofagasta/Mejillones. However, the increases in relative risks are far too great to result from bias due to in-migration of very high-risk persons born elsewhere. We conclude that the effects are most probably due to arsenic in the water and that, if anything, they are diluted by in-migration of people who were born and grew up elsewhere in Chile.

The study's weakness lies in its reliance on death certificates, even though Chilean mortality records are well documented: Laws require that deaths be registered with the Civilian Registration Service (Servicio de Registro Civil), whereas another branch of government, the National Institute of Statistics (Instituto Nacional de Estadísticas), oversees validation of the generated data. Death certificates are coded according to the standard ICD, and the 1996 World Health Statistics cited Chile as having 100, 100, and 98% of all estimated deaths registered for the years 1991, 1993, and 1994, respectively (World Health Organization 1998). However, although death certificates provide reasonably good data for lung cancer studies, they have known limitations for identi- fying death from chronic respiratory disease (Selikoff and Seidman 1992). This leads one to question whether medical practices in region II might have led to overdiagnosis of chronic res- piratory disease as a cause of death placed on death certificates, particularly deaths from bronchiectasis. However, separating out the findings concerning bronchiectasis from other COPD causes of death was conducted with a clear a priori hypothesis. Although previous mention had been made in the literature of bronchiectasis and arsenic, it was the recent finding of a 10-fold increase in bronchiectasis prevalence in persons with high exposure to arsenic and arsenic-caused skin lesions in West Bengal, India (Guha Mazumder et al. 2005), that led us specifically to evaluate bronchiectasis in this study.

Although smoking is strongly associated with mortality from lung cancer and COPD, confounding due to smoking is unlikely. Smoking is not a strong risk factor for bronchiectasis and so would not confound our findings regarding this disease (Barker 2002). And even in extreme form, confounding could not produce the marked elevation of lung cancer relative risks we have found (Axelson 1980). In addition, smoking data do not indicate higher smoking rates in region II than in the rest of Chile, according to a national survey conducted in 1990 (Ministerio de Planificación y Cooperación Nacional Republica de Chile 1992). The survey included the two largest cities in region II (Antofagasta and Calama), which constitute 80% of the region II population; the proportion of smokers in these two cities was found to be lower than the rest of Chile, and the two cities also had a smaller proportion of people who smoked more than one pack per day (Table 2) (Smith et al. 1998). Although there is some evidence that exposure of children to passive smoking in their homes increases the risk of adult lung cancer (Lee et al. 2000; Vineis et al. 2005), an earlier meta-analysis estimated the relative risk to be 0.91 (95% CI, 0.8–1.05) (Boffetta et al. 2000). Even if passive smoking does increase the risk of adult lung cancer, such exposure occurs throughout Chile. Finally, occupational exposures to arsenic, such as in the mining and refining of copper, could contribute to COPD and lung cancer mortality, but these occupational exposures mainly involve men, and our study found similar increases in mortality in both men and women.

Bibliography

[1] Axelson O. 1980. Aspects of confounding and effect modification in the assessment of occupational cancer risk. J Toxicol Environ Health 6(5–6):1127–1131.

[2] Barker A. 2002. Bronchiectasis. N Engl J Med 346:1383–1393.

[3] Boffetta P, Tredaniel J, Greco A. 2000. Risk of childhood cancer and adult lung cancer after childhood exposure to passive smoke: a meta-analysis. Environ Health Perspect 108:73–82.

[4] Chattopadhyay S, Bhaumik S, Nag Chaudhury A, Das Gupta S. 2002. Arsenic induced changes in growth development and apoptosis in neonatal and adult brain cell

[5] Chen CJ, Chuang YC, Lin TM, Wu HY. 1985. Malignant neo- plasms among residents of a blackfoot disease-endemic area in Taiwan: high-arsenic artesian well water and cancers. Cancer Res 45(11 pt 2):5895–5899.

[6] Chen CJ, Wang CJ. 1990. Ecological correlation between arsenic level in well water and age-adjusted mortality from malignant neoplasms. Cancer Res 50(17):5470–5474.

[7] Chen CJ, Wu MM, Lee SS, Wang JD, Cheng SH, Wu HY. 1988. Atherogenicity and carcinogenicity of high-arsenic artesian well water. Multiple risk factors and related malignant neoplasms of blackfoot disease. Arteriosclerosis 8(5):452–460.

[8] Concha G, Vogler G, Lezcano D, Nermell B, Vahter M. 1998. Exposure to inorganic arsenic metabolites during early human development. Toxicol Sci 44(2):185–190.

[9] De BK, Majumdar D, Sen S, Guru S, Kundu S. 2004. Pulmonary involvement in chronic arsenic poisoning from drinking contaminated ground-water. J Assoc Physicians India 52:395–400.

[10] Ferreccio C, Gonzalez CA, Milosavjlevic V, Marshall G, Sancha AM, Smith AH. 2000. Lung cancer and arsenic concentra- tions in drinking water in Chile. Epidemiology 11(6):673–679.

[11] Guha Mazumder DN, Haque R, Ghosh N, De BK, Santra A, Chakraborti D, et al. 2000. Arsenic in drinking water and the prevalence of respiratory effects in West Bengal, India. Int J Epidemiol 29(6):1047–1052.

[12] Guha Mazumder DN, Steinmaus C, Bhattacharya P, von Ehrenstein OS, Ghosh N, Gotway M, et al. 2005. Bronchiectasis in persons with skin lesions resulting from arsenic in drinking water. Epidemiology 16(6):760–765.

[13] Hanlon DP, Ferm VH. 1987. The concentration and chemical status of arsenic in the early placentas of arsenate-dosed hamsters. Environ Res 42(2):546–552.

[14] Hopenhayn C, Ferreccio C, Browning SR, Huang B, Peralta C, Gibb H, et al. 2003. Arsenic exposure from drinking water and birth weight. Epidemiology 14(5):593–602.

[15] Hopenhayn-Rich C, Biggs ML, Smith AH. 1998. Lung and kidney cancer mortality associated with arsenic in drinking water in Córdoba, Argentina. Int J Epidemiol 27(4):561–569.

[16] Hopenhayn-Rich C, Browning SR, Hertz-Picciotto I, Ferreccio C, Peralta C, Gibb H. 2000. Chronic arsenic exposure and risk of infant mortality in two areas of Chile. Environ Health Perspect 108:667–673.

[17] Hopenhayn-Rich C, Hertz-Picciotto I, Browning SR, Ferreccio C, Peralta C. 1999. Reproductive and developmental effects associated with chronic arsenic exposure. In: Arsenic Exposure and Health Effects (Abernathy CO, Calderon RL, Chappell WR, eds). New York:Chapman & Hall, 151–164.

[18] IARC (International Agency for Research on Cancer). 2002. Some Drinking-water Disinfectants and Contaminants, Including Arsenic. IARC Monogr Eval Carcinogen Risks Hum 84:1–19.

[19] Lee CH, Ko YC, Goggins W, Huang JJ, Huang MS, Kao EL, et al. 2000. Lifetime environmental exposure to tobacco smoke and primary lung cancer of non-smoking Taiwanese women. Int J Epidemiol 29(2):224–231.

[20] Milton AH, Rahman M. 2002. Respiratory effects and arsenic contaminated well water in Bangladesh. Int J Environ Health Res 12(2):175–179.

[21] Milton AH, Smith WP, Rahman B, Hasan Z, Kulsum Z, Kulsum U, et al. 2005. Chronic arsenic exposure and adverse pregnancy outcomes in Bangladesh. Epidemiology 16(1):82–86.

[22] Ministerio de Planificación y Cooperación Nacional Republica de Chile. 1992. Illa Encuesta Caracterización Socio Económica Nacional. Santiago, Chile:Ministerio de Planificacion y Cooperacion Nacional Republica de Chile.

[23] NRC (National Research Council). 2001. Arsenic in Drinking Water: 2001 Update. Washington, DC:National Academy Press.

[24] Selikoff IJ, Seidman H. 1992. Use of death certificates in epidemi- ological studies, including occupational hazards: variations in discordance of different asbestos-associated diseases on best evidence ascertainment. Am J Ind Med 22(4):481–492.

[25] Selvin S. 1995. Practical Biostatistical Methods. Belmont, CA:Duxbury Press. Shimizu Y, Schull WJ, Kato H. 1990. Cancer risk among atomic bomb survivors. The RERF Life Span Study. Radiation Effects Research Foundation. JAMA 264(5):601–604.

[26] Smith AH, Goycolea M, Haque R, Biggs ML. 1998. Marked increase in bladder and lung cancer mortality in a region of northern Chile due to arsenic in drinking water. Am J Epidemiol 147(7):660–669.

[27] Tsuda T, Babazono A, Yamamoto E, Kurumatani N, Mino Y, Ogawa T, et al. 1995. Ingested arsenic and internal cancer: a historical cohort study followed for 33 years. Am J Epidemiol 141(3):198–209.

[28] Tsuda T, Nagira T, Yamamoto M, Kurumatani N, Hotta N, Harada M, et al. 1989. Malignant neoplasms among residents who drank well water contaminated by arsenic from a king's yellow factory. J UOEH 11(suppl):289–301.

[29] Vahter M. 1994. Species differences in the metabolism of arsenic compounds. Appl Organomet Chem 8:175–182.

[30] Vineis P, Airoldi L, Veglia P, Olgiati L, Pastorelli R, Autrup H, et al. 2005. Environmental tobacco smoke and risk of respiratory cancer and chronic obstructive pulmonary disease in former smokers and never smokers in the EPIC prospective study. BMJ 330(7486):277–281.

[31] von Ehrenstein OS, Guha Mazumder DN, Hira-Smith M, Ghosh N, Yuan Y, Windham G, et al. 2006. Pregnancy out- comes, infant mortality and arsenic in drinking water in West Bengal, India. Am J Epidemiol 163(7):662–669.

[32] von Ehrenstein OS, Guha Mazumder DN, Yuan Y, Samanta S, Balmes J, Sil A, et al. 2005. Decrements in lung function related to arsenic in drinking water in West Bengal, India. Am J Epidemiol 162(6):533–541.

[33] Waalkes MP, Keefer LK, Diwan BA. 2000. Induction of proliferative lesions of the uterus, testes, and liver in Swiss mice given repeated injections of sodium arsenate: possible estrogenic mode of action. Toxicol Appl Pharmacol 166(1):24–35.

[34] World Health Organization. 1978. International Classification of Diseases, 9th Revision. Geneva:World Health Organization. World Health Organization. 1998. World Health Statistics. Annual 1996. Geneva:World Health Organization. Available: http://www.who.int/whr/1996/en/index.html [accessed 7 September 2005].

[35] Wu MM, Kuo TL, Hwang YH, Chen CJ. 1989. Dose-response rela- tion between arsenic concentration in well water and mor- tality from cancers and vascular diseases. Am J Epidemiol 130(6):1123–1132.

[36] Yoshimoto Y, Kato H, Schull WJ. 1988. Risk of cancer among children exposed in utero to A-bomb radiations, 1950–84. Lancet 2(8612):665–669.

[37] Zaldivar R. 1974. Arsenic contamination of drinking water and foodstuffs causing endemic chronic poisoning. Beitr Pathol 151(4):384–400.

[38] Zaldivar R. 1977. Ecological investigations on arsenic dietary intake and endemic chronic poisoning in man: dose- response curve. Zentralbl Bakteriol [B] 164(5–6):481–484.

[39] Zaldivar R. 1980. A morbid condition involving cardio-vascular, broncho-pulmonary, digestive and neural lesions in children and young adults after dietary arsenic exposure. Zentralbl Bakteriol [B] 170(1–2):44–56.

[40] Zaldivar R, Ghai GL. 1980. Clinical epidemiological studies on endemic chronic arsenic poisoning in children and adults, including observations on children with high- and low-intake of dietary arsenic. Zentralbl Bakteriol [B] 170(5–6):409–421.

Article 3.3

Acute Myocardial Infarction Mortality in Comparison with Lung and Bladder Cancer Mortality in Arsenic-exposed Region II of Chile from 1950 to 2000

Yan Yuan, Guillermo Marshall1, Catterina Ferreccio, Craig Steinmaus1, Steve Selvin, Jane Liaw, Michael N. Bates, and Allan H. Smith

University of California, Berkeley, CA and Universidad Católica de Chile

Abstract. Arsenic in drinking water is known to be a cause of lung, bladder, and skin cancer, and some studies report cardiovascular disease effects. The authors investigated mortality from 1950 to 2000 in the arsenic-exposed region II of Chile (population: 477,000 in 2000) in comparison with the unexposed region V. Increased risks were found for acute myocardial infarction (AMI), with mortality rate ratios of 1.48 for men (95% confidence interval (CI): 1.37, 1.59; $p < 0.001$) and 1.26 for women (95% CI: 1.14, 1.40; $p < 0.001$) during the high-exposure period in region II from 1958 to 1970. The highest rate ratios were for young adult men aged 30–49 years who were born during the high-exposure period with probable exposure in utero and in early childhood (rate ratio¼ 3.23, 95% CI: 2.79, 3.75; $p <$ 0.001). Compared with lung and bladder cancer, AMI mortality was the predominant cause of excess deaths during and immediately after the high-exposure period. Ten years after reduction of exposures, AMI mortality had decreased, and longer latency excess deaths from lung and bladder cancer predominated. With these three causes of death combined, increased mortality peaked in 1991–1995, with estimated excess deaths related to arsenic exposure constituting 10.9% of all deaths among men and 4.0% among women. [1]

Keywords: arsenic; Chile; lung neoplasms; mortality; myocardial infarction; urinary bladder neoplasms; water

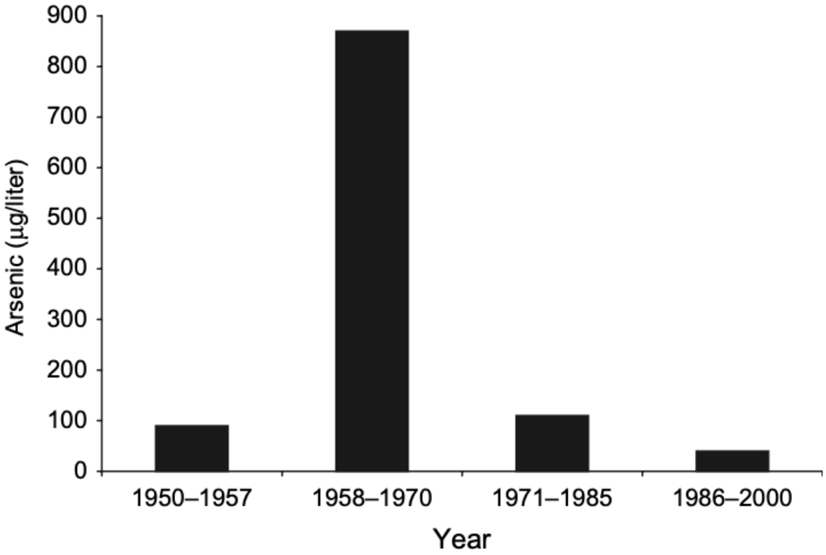
Arsenic in drinking water is classified as carcinogenic to humans by the International Agency for Research on Cancer on the basis of evidence that it causes skin, lung, and bladder cancer (1). Chronic exposure to arsenic in drinking water is also linked to an increased risk of various noncancer health outcomes including dermal, reproductive, pulmonary, and neurologic effects (2, 3). The first evidence of cardiovascular disease associated with arsenic in drinking water came from Antofagasta in region II of Chile, with a case series of 17 deaths from myocardial infarction reported in subjects under the age of 40 years (4). Later evidence mainly from Taiwan suggested that long-term arsenic ingestion may be associated with increased risk of circulatory disease mortality, including cardiovascular disease (5, 6), ischemic heart disease (7–12), cerebrovascular disease (10, 13), and diseases of the arteries, arterioles, and capillaries (14).

Region II of Chile, which had a population of 477,000 in 2000, is unique in the world for investigating the long-term health effects of arsenic in drinking water. One reason is that almost all drinking water in this region is supplied by a few large municipal water sources, with known

[1] Yuan Y, Marshall G, Ferreccio C, Steinmaus C, Selvin S, Liaw J, Bates MN, and Smith AH. 2007. Acute myocardial infarction mortality in comparison with lung and bladder cancer mortality in arsenic-exposed region II of Chile from 1950 to 2000. American Journal of Epidemiology 166 (12), 1381-1391

arsenic concentrations for the past 50 years. This is in contrast to other countries with high arsenic exposures, such as Argentina, Bangladesh, China, India, Taiwan, and the United States, where high arsenic exposures come primarily from wells. Assessing exposure in these areas is extremely difficult because of the large number of wells, the high variability in arsenic concentrations from well to well, and the general lack of historical arsenic measurements.

Fig. 1. Arsenic concentrations (lg/liter) in drinking water before and after an arsenic removal plant was installed in 1971 for Antofagasta and Mejillones (region II), Chile, 1950–2000.



The second unique aspect of arsenic exposure in region II is that it involves a large population with a rapid onset of very high arsenic exposure when rivers contaminated with arsenic began to be used. Exposure was later sharply reduced around 1971 when installation of water treatment plants began. This situation has not been seen before, will probably never recur, and offers an important opportunity to study the health impacts of arsenic. More than half of region II's population live in Antofagasta and Mejillones (current population: 318,000) and were exposed to levels of arsenic greater than 850 lg/liter for a 13-year period (1958–1970) (figure 1).

We recently reported on lung and bladder cancer mortality in region II from 1950 to 2000 (15). We showed that mortality rates from these cancers started to increase about 10 years after the high exposures commenced and did not peak until about 20 years after the start of reductions in exposure. The purpose of this study was to investigate circulatory disease mortality in region II before, during, and after the period of very high arsenic exposure and to compare this mortality with that from lung and bladder cancer, the two major established causes of mortality from arsenic in drinking water. The a priori hypothesis was that circulatory disease mortality might be increased with arsenic exposure, in particular, mortality from acute myocardial infarction (AMI). Our aim was to investigate the latency patterns between onset and decline of exposure and increased circulatory disease mortality. We also planned to assess separately those with in utero and early childhood exposure and those exposed only as adults.

1 Material and Methods

Exposure data

Details concerning the arsenic concentrations in water in region II have been reported previously (16–18). As shown in figure 1, prior to 1958, the drinking water supply in the major city of Antofagasta had an arsenic concentration of about 90 μg/liter. A growing population and increased need for water led to supplementation of Antofagasta's water supply with water from the Toconce and the Holajar rivers that had arsenic concentrations of 800 lg/liter and 1,300 μg/ liter, respectively. The concentration of arsenic in Antofa- gasta's drinking water, along with that of Mejillones which shared the same supply, increased in 1958 to an average of 870 lg/liter. About 90 percent of the population of region II lives in cities and towns. The other towns in the region, with the exception of Taltal, also had high concentrations of arsenic in their drinking water for various overlapping periods. The population-weighted average arsenic concentration in drinking water for the entire region was about 580 μg/liter for about 13 years from 1958 to 1970 (16). With the introduction in 1971 of a water treatment plant, Antofagasta's water arsenic concentration dropped to about 110 μg/liter, and further reductions occurred as a result of treatment plant improvements. In recent years, Antofagasta's water contained about 40 μg of arsenic per liter (16, 17), and the concentration is now below 10 μg/liter, which is the World Health Organization guideline for arsenic in drinking water. Other cities and towns also implemented water treatment strategies or used alternative sources to reduce arsenic levels. By the late 1980s, almost all of the towns with populations over 1,000 had water arsenic concentrations of less than 100 μg/liter. The exception was San Pedro de Atacama (population: 3,700), which has only recently had an arsenic removal plant installed. In contrast, water sources in the rest of Chile have had low levels of arsenic, generally less than 10 μg/liter. The major city of the comparison region V population, Valparaíso, has water concentrations close to 1 μg/ liter (18).

Selection of the comparison population

Electronically stored mortality data were not available for Chile from 1950 to 1970. It was impractical and prohibitively expensive for the study team to nosologize all of the death certificates for Chile for these years. Because of this, it was necessary to select an alternative referent population to span the whole study period, 1950–2000. It was desirable that the referent population be significantly larger than that of region II, in order to maximize statistical precision. After careful consideration, region V, with a population about four times that of region II, was selected. In 1980, the population of region II was 314,807, while the population of region V was 1,230,498. This ratio has been similar throughout the study period of 1950–2000.

Table 1. Numbers of circulatory disease deaths and person-years at risk for adults aged 20 years or above in region II (exposed) and region V (unexposed), Chile, 1950–2000

	Mortality 1950–1957		Mortality 1958–1970		Mortality 1971–1985		Mortality 1986–2000	
	Reg. II	Reg. V	Reg. II	Reg. V	Reg. II	Reg. V	Reg. II	Reg. V
Male person-years	472,234	1,379,972	824,258	2,760,739	1,218,962	4,253,502	1,938,205	6,297,084
Male deaths								
All circulatory diseases	1,803	7,052	3,656	13,467	4,362	18,338	4,604	21,425
Hypertensive disease	243	760	295	963	171	717	230	1,384
Ischemic heart disease	442	1,638	1,496	4,459	2,090	7,228	2,360	9,361
Acute myocardial infarction	321	1,253	1,352	3,618	1,592	5,204	1,736	6,686
Cerebrovascular disease	372	1,589	897	3,805	1,121	5,720	1,153	6,386
Cerebral hemorrhage	274	1,242	409	1,714	300	1,391	381	1,812
Cerebral infarction	67	250	445	1,796	658	3,443	599	3,520
Diseases of arteries, arterioles, and capillaries	208	865	461	1,703	372	1,466	268	965
Female person-years	383,571	1,524,899	755,882	3,106,392	1,201,699	4,805,893	1,892,032	7,034,334
Female deaths								
All circulatory diseases	1,375	6,947	2,508	13,673	3,540	19,003	4,082	21,762
Hypertensive disease	260	898	312	1,311	182	948	287	1,859
Ischemic heart disease	205	1,037	700	3,268	1,304	6,079	1,597	8,228
Acute myocardial infarction	141	770	576	2,332	804	3,608	862	5,185
Cerebrovascular disease	255	1,830	727	4,415	1,132	7,061	1,298	7,369
Cerebral hemorrhage	181	1,424	324	1,967	238	1,688	376	1,876
Cerebral infarction	61	304	373	2,082	744	4,151	702	4,244
Diseases of arteries, arterioles, and capillaries	168	930	329	1,882	363	1,863	260	1,021

To ensure that region V was an appropriate choice, preliminary investigations were conducted to compare per capita income, smoking rates, and death certification among region II, region V, and national data for the whole of Chile. Per capita income in region V in 1990 was similar to that of the rest of the country (US $2,053vs.US$2,011). Region II had higher per capita income (US \$3,853), but this was the result of exports generated by the mining industry rather than signifying higher personal income. Smoking surveys were carried out on random population samples in 1990 and 1992, both years giving similar data. In 1990, 26.6 percent of men and 19.3 percent of women in Chile said they smoked. The corresponding percentages from regions II and V were similar: 27.4 percent and 28.5 percent for men and 16.6 percent and 20.2 percent for women, respectively (19). We also obtained information concerning death certification by health services regions in the country from a study conducted in 1983 (20). For the whole country, 85.6 percent of the death certificates in that year were certified by a physician. The corresponding percentages in regions II and V were 89.8 percent and 94.5 percent. Thus, the large majority of death certificates were completed by physicians, with both region II and region V having a higher percentage than the national average. The information above gave assurance that region V was a suitable referent population for two major determinants of cardiovascular morality, socioeconomic status and cigarette smoking, and that it was also a suitable referent population based on quality of death certification.

1.1 Mortality data collection

For the years 1950–1970, all the death certificates for region II and region V were photographed, displayed on computer monitors, and coded by trained nosologists according to the International Classification of Diseases, Ninth Revision (ICD-9). Death certificates from both regions were intermingled, and nosologists were kept blind as to the region from which each death certificate originated. Computerized mortality data first became available in Chile in 1971. These data, already coded to ICD-9 for all regions of Chile for the years 1971–1979 (excluding 1976), were obtained from the Chilean National Institute of Statistics (Instituto Nacional Estadisticas). For 1976, the information that is normally stored on computer disk at the Institute was never completed because of political unrest in the country. Mortality data for all regions of Chile for the years 1980–2000 were obtained from the Ministry of Health. ICD-9 codes had been used for 1980–1998, and International Classification of Diseases, Tenth Revision (ICD-10), codes had been used for 1999 and 2000. These codes were used to group cause-specific mortality into diseases of the circulatory system (ICD-9 codes 390–459; ICD-10 codes I00–I99); hypertensive disease (ICD-9 codes 401–405; ICD-10 codes I10–I15); ischemic heart disease (ICD-9 codes 410–414; ICD-10 codes I20–I25); AMI (ICD-9 code 410; ICD-10 codes I21–I22); cerebrovascular disease (ICD-9 codes 430–438; ICD-10 codes I60–I69); subarachnoid and intracerebral hemorrhage (ICD-9 codes 430–432; ICD-10 codes I60–I62); cerebral infarction (ICD-9 codes 434 and 436; ICD-10 code I63); and diseases of the arteries, arterioles, and capillaries (ICD-9 codes 440–448; ICD-10 codes I70–I78).

Annual estimates of the population living in regions II and V stratified by age and gender for the period 1950–2000 were obtained from the National Institute of Statistics. The estimates were obtained by linear interpolation between census data collected approximately every 10 years.

Table 2. Age-adjusted rate ratios for circulatory disease mortality for region II (exposed) compared with region V (unexposed), Chile, for 1950-1957 (before the peak exposures in region II), 1958-1970 (the peak exposure period), 1971-1985 (after an arsenic removal plant was installed in the major city of Antofagasta), and 1986-2000

	Mortality, 1950-1957			Mortality, 1958-1970			Mortality, 1971-1985			Mortality, 1986-2000		
	Rate ratio	95% CI	p value	Rate ratio	95% CI	p value	Rate ratio	95% CI	p value	Rate ratio	95% CI	p value
Males												
All circulatory diseases	0.90	(0.82, 0.99)	0.03	1.09	(1.06, 1.12)	<0.001	1.11	(1.06, 1.16)	<0.001	1.03	(0.98, 1.08)	0.28
Hypertensive disease	1.15	(0.98, 1.35)	0.08	1.25	(1.06, 1.46)	<0.01	1.12	(0.98, 1.28)	0.08	0.84	(0.70, 1.00)	0.06
Ischemic heart disease	0.93	(0.81, 1.07)	0.31	1.34	(1.22, 1.47)	<0.001	1.35	(1.24, 1.48)	<0.001	1.21	(1.05, 1.39)	<0.01
Acute myocardial infarction	0.88	(0.74, 1.03)	0.11	1.48	(1.37, 1.59)	<0.001	1.41	(1.27, 1.56)	<0.001	1.21	(1.00, 1.47)	0.05
Cerebrovascular disease	0.81	(0.68, 0.97)	0.02	0.95	(0.87, 1.03)	0.22	0.91	(0.81, 1.03)	0.13	0.86	(0.81, 0.92)	<0.001
Cerebral hemorrhage	0.75	(0.65, 0.88)	<0.001	0.94	(0.84, 1.05)	0.25	0.94	(0.84, 1.05)	0.28	0.90	(0.78, 1.04)	0.16
Cerebral infarction	0.97	(0.73, 1.30)	0.86	1.01	(0.90, 1.13)	0.87	0.91	(0.76, 1.07)	0.25	0.85	(0.78, 0.91)	<0.001
Diseases of arteries, arterioles, and capillaries	0.94	(0.73, 1.21)	0.64	1.14	(0.96, 1.36)	0.14	1.25	(1.04, 1.49)	0.02	1.40	(1.07, 1.83)	0.01
Females												
All circulatory diseases	0.96	(0.86, 1.07)	0.50	0.94	(0.90, 0.99)	0.99	1.01	(0.97, 1.05)	0.61	1.01	(0.95, 1.08)	0.73
Hypertensive disease	1.42	(1.27, 1.59)	<0.001	1.24	(1.15, 1.33)	<0.001	1.05	(0.93, 1.20)	0.42	0.85	(0.69, 1.03)	0.10
Ischemic heart disease	0.96	(0.73, 1.25)	0.75	1.11	(0.99, 1.24)	0.08	1.18	(1.12, 1.24)	<0.001	1.06	(0.93, 1.22)	0.37
Acute myocardial infarction	0.88	(0.67, 1.14)	0.33	1.26	(1.14, 1.40)	<0.001	1.21	(1.11, 1.31)	<0.001	0.90	(0.71, 1.14)	0.37
Cerebrovascular disease	0.67	(0.61, 0.73)	<0.001	0.84	(0.79, 0.90)	<0.001	0.86	(0.81, 0.92)	<0.001	0.94	(0.86, 1.03)	0.16
Cerebral hemorrhage	0.60	(0.55, 0.66)	<0.001	0.82	(0.74, 0.92)	<0.01	0.72	(0.64, 0.81)	<0.001	0.98	(0.89, 1.08)	0.72
Cerebral infarction	0.99	(0.78, 1.27)	0.95	0.93	(0.86, 1.01)	0.10	0.97	(0.91, 1.04)	0.46	0.90	(0.80, 1.02)	0.10
Diseases of arteries, arterioles, and capillaries	0.95	(0.80, 1.14)	0.54	0.95	(0.84, 1.07)	0.36	1.10	(0.91, 1.33)	0.33	1.42	(1.19, 1.69)	<0.001

CI = Confidence Interval

1.2 Statistical analysis

To investigate the temporal relation of mortality to changes in the concentration of arsenic, we categorized cal- endar years into four time periods on the basis of period of high exposure in region II: 1950–1957 (preexposure), 1958–1970 (high exposure), 1971–1985 (intermediate exposure), and 1986–2000 (low exposure). We estimated mortality rate ratios using Poisson regression analysis for each cause of death, comparing region II with region V in each exposure time period for men and women separately and age adjusted in 10-year age strata from 20 to greater than 80 years.

Poisson regression analysis was performed using the PROC GENMOD procedure provided in SAS, version 8.2, software (SAS Institute, Inc., Cary, North Carolina). Anal- yses were conducted with the link function as the log and the offset as the log of the total population in each region, sex, and age stratum. To further help identify the trends in mortality rate ratios, we calculated and plotted 5-year Poisson regression rate ratios for the entire study period, 1950–2000. A separate analysis was done to evaluate the impacts of early life arsenic exposure on the risks of AMI mortality later in life. To do this, we defined two birth cohorts based on the high-exposure period in region II (1958–1970): those born during the high-exposure period and those born in 1950–1957, just before the high-exposure period. Those born during the high-exposure period would have experienced exposure in utero, as well as early childhood, while those born just before 1958 would have experienced high exposure during childhood but not in utero. In this analysis, we focused on the age group 30–49 years, since all of these subjects would be aged 50 years or younger by the end of our study period (the year 2000). For the years 1989–2000, deaths and population estimates were available for two major cities in region II, Antofagasta and Mejillones, which had the highest arsenic exposure. We therefore were able to compare mortality rates in Antofagasta and Mejillones with those of region V, using Poisson regression estimation of rate ratios, and also with the rest of Chile using standardized mortality ratios, since computerized mortality data were available for the whole county for these years.

The numbers of excess deaths due to AMI, lung cancer, and bladder cancer in region II for the years 1950–2000 were estimated. We grouped 1950–1957 because it was the preexposure period. The high-exposure period 1958– 1970 was divided roughly in half into two periods, 1958–1964 and 1965–1970. We first estimated rate ratios for AMI, lung cancer, and bladder cancer, comparing region II with region V in each grouped time period for men and women separately, and age adjusted in 10-year age strata from age 20 to age 80 years or more using Poisson regression analysis. Then, the estimated numbers of excess deaths for each cause of death were calculated for each grouped time period: (rate ratio: $RR - 1)/RR) \times N$, where N is the total number of deaths from the cause of death in that time period in region II.

2 Results

Table 1 presents the numbers of the circulatory disease deaths and person-years at risk. The age-adjusted mortality rate ratios comparing region II with region V are shown in table 2 for men and women separately. The mortality rates for AMI in region II were increased during the high-exposure period 1958–1970 (men: RR¼ 1.48, 95 percent confidence interval (CI): 1.37, 1.59 (p ¡ 0.001); women: RR¼ 1.26, 95 percent CI: 1.14, 1.40 (p ¡ 0.001)). Region II rates re- mained elevated during the immediate postexposure period 1971–1985 (men: RR¼ 1.41, 95 percent CI: 1.27, 1.56 (p ¡ 0.001); women: RR¼ 1.21, 95 percent CI: 1.11, 1.31 (p ¡ 0.001)) and then gradually decreased during the final study period 1986–2000 (men: RR¼ 1.21, 95 percent CI: 1.00, 1.47 (p = 0.05); women: RR¼ 0.90, 95 percent CI: 0.71, 1.14 (p = 0.37)). Figure 2 presents the time pattern using 5-year mortality rate ratio estimates.

Fig. 2. Age-adjusted mortality rate ratios and 95intervals for acute myocardial infarction for males and females, region II (exposed) compared with region V (unexposed), Chile, 1950–2000. Each point represents an estimate for 5 years and is plotted at the midpoint of the 5-year period, starting with the estimate for 1950– 1954, which is plotted at the year 1952. (Note: The years 1950–1957 were prior to exposure, followed by high exposure from 1958 to 1970. During 1971–1985, there was intermediate exposure, and in 1986– 2000, low exposure.)

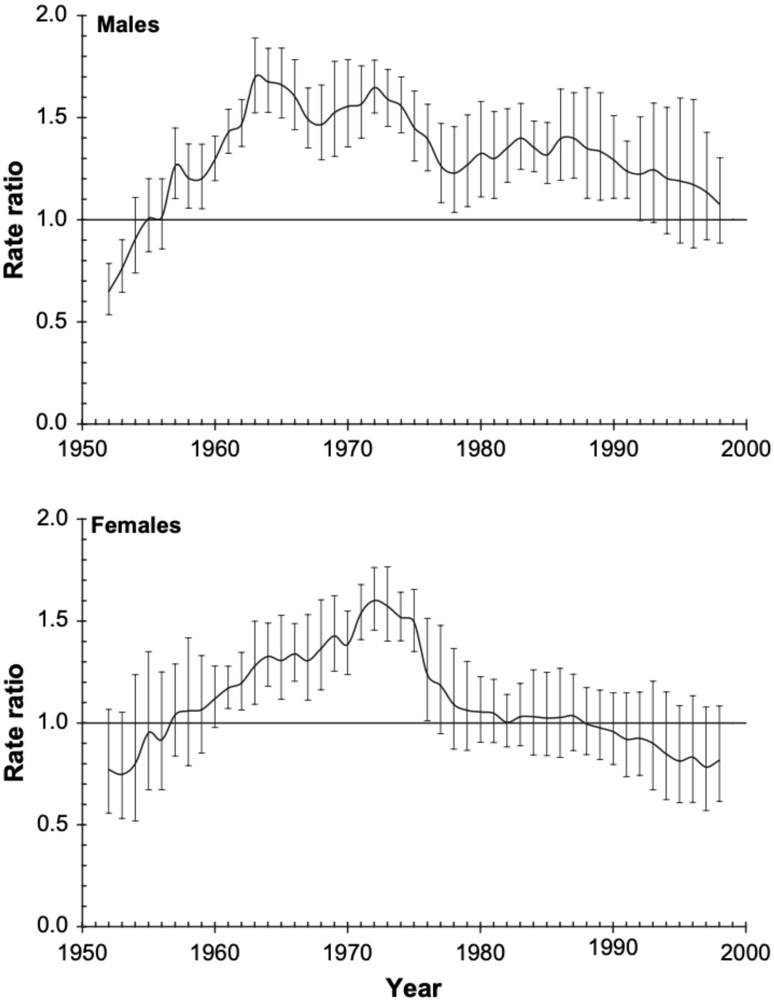


Table 3. Numbers of deaths from acute myocardial infarction and rate ratios according to sex and age for region II (exposed) compared with region V (unexposed), Chile, for 1950-1957 (before the peak exposures in region II), 1958-1970 (the peak exposure period), 1971-1985 (after an arsenic removal plant was installed in the major city of Antofagasta), and 1986-2000

	Mortality, 1950-1957					Mortality, 1958-1970					Mortality, 1971-1985					Mortality, 1986-2000				
	II deaths	V deaths	Rate ratio	95% CI	p value	II deaths	V deaths	Rate ratio	95% CI	p value	II deaths	V deaths	Rate ratio	95% CI	p value	II deaths	V deaths	Rate ratio	95% CI	p value
Males (years)																				
20-29	6	11	1.43	(0.53, 3.86)	0.48	10	22	1.41	(0.67, 2.98)	0.36	10	16	1.98	(0.90, 4.35)	0.09	8	17	1.42	(0.61, 3.30)	0.41
30-39	14	50	0.80	(0.44, 1.44)	0.46	57	77	2.27	(1.61, 3.20)	<0.001	43	65	2.08	(1.42, 3.06)	<0.001	49	65	2.14	(1.48, 3.10)	<0.001
40-49	59	162	1.13	(0.84, 1.52)	0.42	151	313	1.65	(1.36, 2.00)	<0.001	139	297	1.61	(1.31, 1.97)	<0.001	157	229	2.07	(1.69, 2.54)	<0.001
50-59	80	282	0.89	(0.70, 1.15)	0.37	297	758	1.47	(1.29, 1.68)	<0.001	337	804	1.66	(1.46, 1.89)	<0.001	339	839	1.46	(1.29, 1.66)	<0.001
60-69	91	399	0.74	(0.59, 0.93)	0.01	406	1,150	1.41	(1.26, 1.58)	<0.001	435	1,455	1.37	(1.24, 1.53)	<0.001	497	1,735	1.27	(1.15, 1.41)	<0.001
70-79	64	256	1.01	(0.77, 1.33)	0.93	328	934	1.47	(1.30, 1.67)	<0.001	398	1,666	1.24	(1.11, 1.39)	<0.001	465	2,177	1.12	(1.02, 1.24)	0.02
>80	7	93	0.43	(0.20, 0.92)	0.03	103	364	1.35	(1.09, 1.69)	<0.01	230	901	1.31	(1.14, 1.52)	<0.001	221	1,624	0.80	(0.70, 0.92)	<0.01
Females (years)																				
20-29	5	10	1.77	(0.60, 5.17)	0.30	6	16	1.37	(0.53, 3.49)	0.52	1	6	0.58	(0.07, 4.80)	0.61	4	0	-	-	-
30-39	7	33	0.81	(0.36, 1.84)	0.62	20	30	2.51	(1.43, 4.42)	<0.01	12	24	1.80	(0.90, 3.59)	0.10	5	12	1.36	(0.48, 3.87)	0.56
40-49	8	67	0.50	(0.24, 1.03)	0.06	38	135	1.20	(0.84, 1.72)	0.32	31	117	1.07	(0.72, 1.60)	0.72	34	89	1.37	(0.92, 2.03)	0.12
50-59	37	130	1.27	(0.88, 1.84)	0.19	98	339	1.38	(1.10, 1.72)	<0.01	81	316	1.19	(0.93, 1.52)	0.16	79	285	1.16	(0.90, 1.49)	0.25
60-69	31	223	0.65	(0.45, 0.95)	0.02	142	631	1.15	(0.96, 1.38)	0.14	196	772	1.36	(1.17, 1.60)	<0.001	223	846	1.30	(1.12, 1.50)	<0.001
70-79	33	213	0.81	(0.56, 1.17)	0.26	162	693	1.22	(1.03, 1.44)	<0.03	262	1,177	1.24	(1.09, 1.42)	<0.01	247	1,656	0.83	(0.73, 0.95)	<0.01
>80	20	94	1.26	(0.78, 2.04)	0.36	110	488	1.32	(1.07, 1.63)	<0.01	221	1,196	1.08	(0.93, 1.24)	0.31	270	2,297	0.69	(0.61, 0.79)	<0.001

II = Region II, V =Region V, CI = Confidence Interval

Hypertensive disease mortality in men increased during the high-exposure period 1958–1970 (RR = 1.25, 95 percent CI: 1.06, 1.46 (p < 0.01)) and then gradually decreased during the intermediate-exposure period 1971–1985 (RR = 1.12, 95 percent CI: 0.98, 1.28 (p = 0.08)) (table 2). The temporal relation between arsenic exposure and hypertensive disease mortality was not evident in women. Rate ratios for cerebrovascular disease mortality were not increased in region II for any time period. Indeed, the rate ratios for cerebral hemorrhage in region II were reduced in all study periods, especially among women, without evidence of a temporal pattern related to arsenic exposure. Increases in mortality from diseases of the arteries, arterioles, and capillaries (which includes peripheral vascular disease) can be seen among men and women, especially in the final period 1986–2000.

Table 3 presents AMI rate ratios by age group. Leaving aside the age group 20–29 years that involves small numbers, there was a general pattern of an inverse relation of rate ratios with age in the high-exposure and following time periods. The highest rate ratios among both men and women were for the age group 30–39 years.

We next estimated rate ratios comparing Antofagasta and Mejillones with region V for AMI mortality in two birth cohorts, those born in 1950–1957 just before the high- exposure period and those born in 1958–1970 during the high-exposure period, for ages at death of 30–39 and 40–49 years separately and combined and for men and women separately (table 4). Compared with those for region V, the rate ratios for Antofagasta and Mejillones men aged 30–49 years were increased for those born in the period 1950–1957 (RR = 2.56, 95 percent CI: 1.26, 5.18 (p < 0.001)) and those born in the period 1958–1970 (RR = 3.23, 95 percent CI: 2.79, 3.75 (p < 0.001)). Table 4 also shows standardized mortality ratios comparing Antofagasta and Mejillones with the rest of Chile, with findings similar to the comparisons with region V. For men born in 1950–1957, the standardized mortality ratio for those aged 30–49 years was 2.51 compared with a rate ratio of 2.56 for the comparison with region V. For men born in 1958–1970, the standardized mortality ratio was 2.72 compared with a rate ratio of 3.23 for the comparison with region V. The relative risk estimates for females are generally lower than the estimates for men, but it should be noted that the upper limits of the confidence intervals of the estimates for women are very high.

Table 5 shows the estimated numbers of excess deaths from AMI, lung cancer, and bladder cancer in region II for the years 1950–2000. During the study period, 38 percent of the excess deaths in men and 32 percent in women were from AMI. Figure 3 shows the number of those excess deaths divided by all deaths. From 1958 to 1979, the majority of excess deaths among both men and women were from AMI. After 1979, the excess deaths associated with AMI decreased, while those associated with lung cancer and bladder cancer continued to increase and remained elevated up to the year 2000. Table 5 also presents the percentage of all deaths estimated to be excess deaths attributable to arsenic. For both men and women, the peaks were in the period 1991–1995, more than 30 years after the high exposures commenced and about 20 years after the installation of an arsenic removal plant for Antofagasta and Mejillones. The peak for men was 10.9 percent, which means that just over one of 10 deaths among men in the period 1991–1995 is attributable to arsenic in drinking water, assuming a causal relation and no bias. The peak for women was lower, reaching 4 percent of all deaths in 1991–1995, suggesting that one of 25 deaths among women might be attributable to arsenic.

Table 4. Rate ratios * and standardized mortality ratios † for acute myocardial infarction mortality in the years 1989-2000 for children born in 1950-1957 and in 1958-1970, the peak exposure periods in Antofagasta and Mejillones, Chile

	Born in 1950-1957					Born in 1958-1970				
	Region II deaths	Region V deaths	Rate ratio	95% CI	p value	Region II deaths	Region V deaths	Rate ratio	95% CI	p value
Antofagasta and Mejillones (region II) vs. region V										
Males (years)										
30-39	6	28	1.11	(0.46, 2.68)	0.82	18	28	3.32	(1.84, 6.01)	<0.001
40-49	47	81	3.06	(2.14, 4.39)	<0.001	2	4	2.61	(0.48, 14.2)	0.27
Pooled	53	109	2.56	(1.26, 5.18)	<0.001	20	32	3.23	(2.79, 3.75)	<0.001
Females (years)										
30-39	0	5	0	-	-	2	6	1.95	(0.39, 9.68)	0.41
40-49	8	33	1.47	(0.68, 3.18)	0.33	0	0	-	-	-
Pooled	8	38	1.27	(0.55, 2.94)	0.57	2	6	1.95	(0.39, 9.68)	0.41

	Born in 1950-1957					Born in 1958-1970				
	Observed deaths	Expected deaths	SMR	95% CI	p value	Observed deaths	Expected deaths	SMR	95% CI	p value
Antofagasta and Mejillones vs. the rest of Chile										
Males (years)										
30-39	6	3.73	1.61	(0.59, 3.50)	0.17	18	6.11	2.94	(1.74, 4.65)	<0.001
40-49	47	16.7	2.81	(2.06, 3.73)	<0.001	2	1.24	1.62	(0.20, 5.84)	0.35
Pooled	53	21.1	2.51	(1.88, 3.29)	<0.001	20	7.35	2.72	(1.66, 4.20)	<0.001
Females (years)										
30-39	0	0.86	0	-	-	2	1.34	1.49	(0.18, 5.38)	0.39
40-49	8	4.59	1.74	(0.75, 3.43)	0.09	0	0.16	0	-	-
Pooled	8	5.58	1.43	(0.62, 2.83)	0.20	2	1.51	1.33	(0.16, 4.80)	0.44

CI = Confidence Interval

SMR = Standarized mortality ratio

* Rate ratios for Antofagasta and Mejillones compared with region V.

† Standardized mortality ratios for Antofagasta and Mejillones compared with the rest of Chile.

3 Discussion

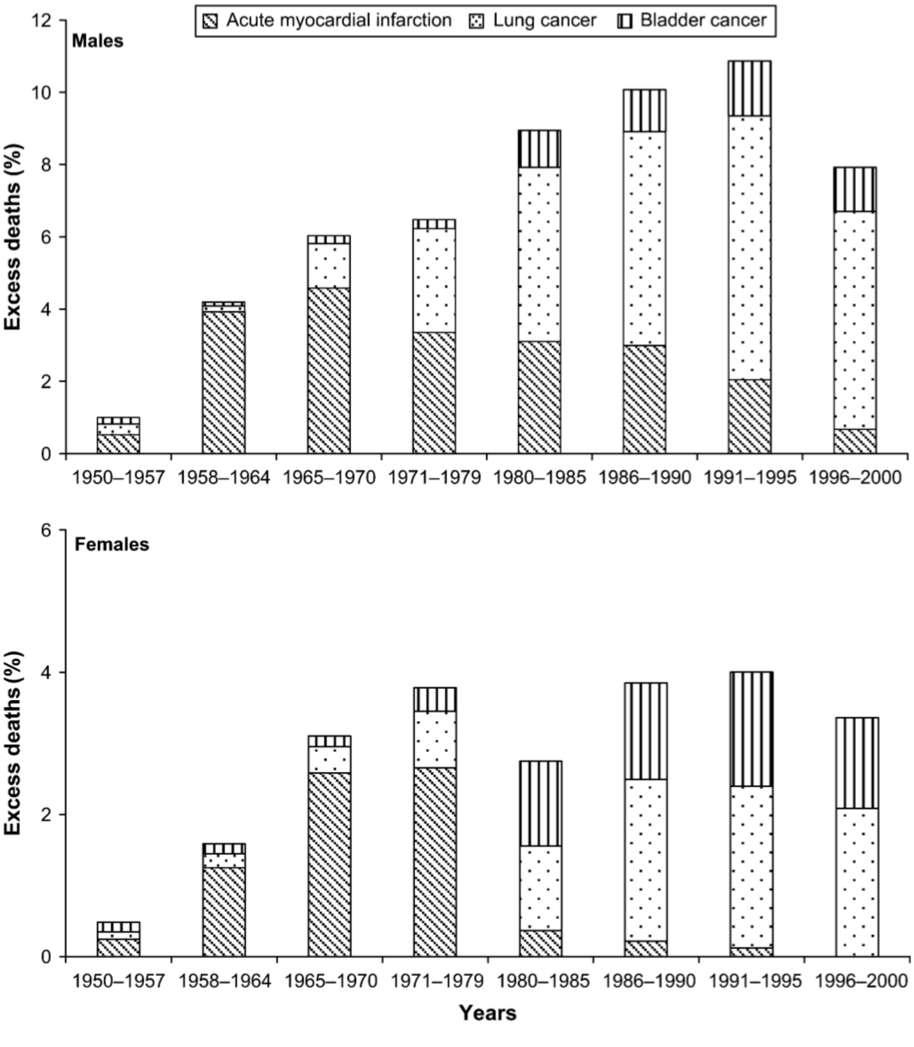
This 50-year mortality study has demonstrated a clear increase in deaths from AMI, with increased risks at the same time that high exposures to arsenic in drinking water started and declining risks about 10 years after exposures were reduced (figure 2). We believe this is an important addition to the body of evidence linking increased mortality from AMI with arsenic in drinking water (4, 7–12), and it is the first study to map out latency from onset of exposure. The study is by far the largest to date on circulatory disease mortality, with more than 8,000 AMI deaths in the exposed population, over 10 times more than those in the largest study in Taiwan. This is also the first study to show that excess AMI deaths predominated during the high-exposure period and for about 10 years thereafter (figure 3). Later, lung cancer and bladder cancer became the predominant contributors to excess deaths. Interestingly, smoking causes both lung and bladder cancer with long latencies, but the risk of AMI associated with smoking increases rapidly (as evidenced by high relative risks in young adults) and also declines rapidly following cessation of smoking (21). However, we have no explanation for the longer continuation of increased risks among men than among women following reduction in arsenic exposure (figure 2).

The main potential limitation of this study is the ecologic design. Potential biases that can result from the lack of individual data on exposure and confounding factors are well known (22). However, we do not believe that ecologic bias is a problem in our study. The first reason for this confidence in our results relates to exposure, since essentially everyone living in region II was exposed to arsenic at high concentrations while those in region V and the rest of Chile were not. This exposure contrast is markedly different from that of most ecologic studies in which it is unclear if the individuals who get disease are those who were actually exposed or not. In our study in Chile, we can state that there were increased rates of AMI in region II (where virtually everyone was exposed to high concentrations of arsenic in drinking water) compared with region V (where all drinking water sources contained low arsenic concentrations). Some people who died in region II would have recently migrated from another part of Chile where they were not exposed, and some from region II may have migrated in the opposite direction. However, the effect of such migration would be to dilute exposure contrasts and would, therefore, reduce, rather than cause, the positive associations we identified. From 1965 to 2000, annual internal migration among regions was only 0.6 percent, compared with 1.2 percent in Argentina, 3.1 percent in the United Kingdom, and 6.6 percent in United States (23). To conclude, although our ecologic study lacks individual data on exposure, the only likely ecologic bias would be a small bias toward the null from migration.

Our study also did not have individual data on confound- ing factors, but for two reasons it is very unlikely that our findings could be due to confounding. The first reason relates to timing. For confounding factors to explain the rise and fall in AMI mortality that we saw, they would have to have a similar relation in time to the rise and fall in arsenic concentrations. For example, if smoking were to explain the increasing and then decreasing mortality rate ratios between regions II and V, there would have to be a sudden rise in smoking rates in region II compared with region V in the 1950s, followed by a return to similar smoking rates in the 1970s.

The second reason for rejecting confounding as an explanation relates to the magnitude of mortality rate ratios identified. The rate ratio estimate for AMI mortality among men in region II was 1.48 (95 percent CI: 1.37, 1.59) during the high-exposure period, compared with region V (table 2). We estimated the magnitude of differences in smoking prevalence between region II and region V that would be needed to produce a mortality rate ratio of 1.48 between the regions using the method first proposed by Axelson (24). We obtained a relative risk estimate for smokers

Fig. 3. Excess deaths as a percentage of total deaths due to acute myocardial infarction, lung cancer, and bladder cancer for males and females, region II (exposed) compared with region V (unexposed), Chile, 1950–2000. (Note: The years 1950–1957 were prior to exposure, followed by high exposure from 1958 to 1970. During 1971–1985, there was intermediate exposure, and in 1986–2000, low exposure.)



from a large cohort study of about 140,000 men that found that the AMI mortality rate ratio was 2.11 for current cigarette smokers who have smoked more than 20 pack-years compared with never smokers (25). Using relative risk estimates for smokers, we estimated that, if 20 percent of men smoked in region V, then at least 70 percent of men in region II would have to smoke to produce a rate ratio of 1.45 or more. If 15 percent of men had smoked in region V, then at least 60 percent would have to have been smokers in region II to result in a population rate ratio of at least 1.43. Such temporary differences in smoking practices are extremely unlikely when the data we have from 1990 show no evidence of any difference in smoking between the regions. Given this example with a strong risk factor such as smoking, it is also unlikely that other confounding factors, including diet and exercise or a combination of them, could produce the magnitude of rate ratios that we found and their trends over time. For all these reasons, we believe that confounding is unlikely and, although the study is ecologic in design, that it provides strong evidence of a causal relation between arsenic in drinking water and AMI mortality.

Table 5. Excess deaths due to acute myocardial infarction, lung cancer, and bladder cancer for males and females, region II (exposed) compared with region V (unexposed), Chile, for the preexposure period 1950-1957, high-exposure period 1958-1970, intermediate-exposure period 1971-1985, and low-exposure period 1986-2000

Years	Total deaths	Excess deaths due to acute myocardial infarction	Excess deaths due to lung cancer	Excess deaths due to bladder cancer	Total excess deaths	Excess deaths as a percentage of total deaths
Males						
1950-1957	5,604	29	17	10	56	1.00
1958-1964	5,650	222	9	6	237	4.19
1965-1970	5,025	230	62	11	303	6.03
1971-1979*	7,966	267	229	20	516	6.48
1980-1985	6,285	195	303	64	562	8.94
1986-1990	5,152	154	305	60	519	10.07
1991-1995	5,639	115	412	86	613	10.87
1996-2000	5,944	40	358	73	471	7.92
Total	47,265	1,252	1,695	330	3,277	6.93
Females						
1950-1957	3,722	9	4	5	18	0.48
1958-1964	3,596	45	7	5	57	1.59
1965-1970	3,251	84	12	5	101	3.11
1971-1979*	5,158	137	41	17	195	3.78
1980-1985	3,998	12	39	39	90	2.75
1986-1990	3,793	8	84	50	142	3.85
1991-1995	4,079	5	92	65	162	4.00
1996-2000	4,568	0	113	69	182	3.36
Total	32,165	300	392	255	947	2.94

* Excluding 1976 data that were not available.

An important finding from this mortality study is the large number of excess deaths attributed to arsenic. In 1991–1995, around 35 years after the highest arsenic exposures commenced, this amounted to about 10 percent of all deaths among men in this period and 4 percent among women. Such high proportions of deaths are unprecedented for any long-term general population environmental exposures. The higher impact on men than women could partly be due to the synergistic effect of arsenic exposure with cigarette smoking, which has already been demonstrated for lung cancer (17).

This study is the first epidemiologic study to report the impact of early life arsenic exposure on mortality from AMI. We recently reported markedly increased mortality from lung cancer and bronchiectasis in the same birth cohort of young adults aged 30–49 years in region II of Chile after probable exposure to arsenic in utero and in early childhood, a finding which we noted provided "some of the first human evidence of effects from environmental exposures to toxic chemicals in utero and early childhood resulting in disease in adults" (26, p. 1296). With the current findings, we can add mortality from AMI in young adults as another possible consequence of early life arsenic exposure.

Evidence of associations between arsenic in water supplies and circulatory disease mortality has previously been found in several high-dose studies from Taiwan (5, 6, 8–13). Previous systematic reviews concerning arsenic in drinking water and cardiovascular disease have been inconclusive as to whether or not there is a causal relation (27, 28). We believe that the clear-cut evidence concerning AMI mortality that has emerged from this mortality study in Chile makes an important addition to our knowledge about arsenic and cardiovascular disease.

Concerning other circulatory disease outcomes, we found no evidence of increased cerebrovascular disease mortality and little evidence of an increase in peripheral vascular disease mortality and hypertensive heart disease mortality. Regarding cerebrovascular disease, findings in Taiwan in relation to arsenic are weaker than the evidence concerning cardiovascular disease mortality (6, 10). Interestingly, Hertz-Picciotto et al. (29) presented evidence from studies of workers inhaling arsenic in the workplace showing that there might be increased mortality from cardiovascular disease and not from cerebrovascular disease.

We conclude that the major impact of arsenic in drinking water on circulatory disease involves AMI and that, in the initial years, it is the main cause of death from arsenic in drinking water, superseded in later years by excess mortality from lung and bladder cancer. Based on the large proportion of excess deaths that we identified, the overall increase in mortality due to arsenic in drinking water in the population of region II of Chile is greater than ever found for mortality from any other environmental exposure in any other population in the world.

Bibliography

[1] International Agency for Research on Cancer. Some drinking- water disinfectants and contaminants, including arsenic. Lyon, France: World Health Organization, 2004.

[2] National Research Council. Arsenic in drinking water: 2001 update. Washington, DC: National Academy Press, 2001.

[3] National Research Council. Arsenic in drinking water. Washington, DC: National Academy Press, 1999.

[4] Zaldivar R. A morbid condition involving cardio-vascular, broncho-pulmonary, digestive and neural lesions in children and young adults after dietary arsenic exposure. Zentralbl Bakteriol [B] 1980;170:44–56.

[5] Chen CJ, Wu MM, Lee SS, et al. Atherogenicity and carci- nogenicity of high-arsenic artesian well water. Multiple risk factors and related malignant neoplasms of blackfoot disease. Arteriosclerosis 1988;8:452–60.

[6] Wu MM, Kuo TL, Hwang YH, et al. Dose-response relation between arsenic concentration in well water and mortality from cancers and vascular diseases. Am J Epidemiol 1989; 130:1123–32.

[7] Tsuda T, Nagira T, Yamamoto M, et al. An epidemiological study on cancer in certified arsenic poisoning patients in Tor- oku. Ind Health 1990;28:53–62.

[8] Chen CJ, Chiou HY, Chiang MH, et al. Dose-response rela- tionship between ischemic heart disease mortality and long- term arsenic exposure. Arterioscler Thromb Vasc Biol 1996; 16:504–10.

[9] Hsueh YM, Wu WL, Huang YL, et al. Low serum carotene level and increased risk of ischemic heart disease related to long-term arsenic exposure. Atherosclerosis 1998;141: 249–57.

[10] Tsai SM, Wang TN, Ko YC. Mortality for certain diseases in areas with high levels of arsenic in drinking water. Arch En- viron Health 1999;54:186–93.

[11] Tseng CH, Chang CK, Tseng CP, et al. Long-term arsenic ex- posure and ischemic heart disease in arseniasis-hyperendemic villages in Taiwan. Toxicol Lett 2003;137:15–21.

[12] Chang CC, Ho SC, Tsai SS, et al. Ischemic heart disease mortality reduction in an arseniasis-endemic area in south- western Taiwan after a switch in the tap-water supply system. J Toxicol Environ Health A 2004;67:1353–61.

[13] Chiou HY, Huang WI, Su CL, et al. Dose-response relation- ship between prevalence of cerebrovascular disease and in- gested inorganic arsenic. Stroke 1997;28:1717–23.

[14] Engel RR, Smith AH. Arsenic in drinking water and mor- tality from vascular disease: an ecologic analysis in 30 counties in the United States. Arch Environ Health 1994; 49:418–27.

[15] Marshall G, Ferreccio C, Yuan Y, et al. Fifty-year study of lung and bladder cancer mortality in Chile related to arsenic in drinking water. J Natl Cancer Inst 2007;99:920–8.

[16] Smith AH, Goycolea M, Haque R, et al. Marked increase in bladder and lung cancer mortality in a region of northern Chile due to arsenic in drinking water. Am J Epidemiol 1998;147: 660–9.

[17] Ferreccio C, Gonzalez CA, Milosavjlevic V, et al. Lung cancer and arsenic concentrations in drinking water in Chile. Epide- miology 2000;11:673–9.

[18] Hopenhayn C, Ferreccio C, Browning SR, et al. Arsenic ex- posure from drinking water and birth weight. Epidemiology 2003;14:593–602.

[19] CASEN. Illa Encuestra CASEN (Caracterizacion Socio Economica Nacional). Santiago, Chile: Ministerio de Planificacion y Cooperacion Nacional Republica de Chile, 1992.

[20] Castillo B, Mardones G. Medical certification of deaths in the health services of Chile. (In Spanish). Rev Med Chil 1986; 114:693–700.

[21] Teo KK, Ounpuu S, Hawken S, et al. Tobacco use and risk of myocardial infarction in 52 countries in the INTERHEART study: a case-control study. Lancet 2006;368:647–58.

[22] Morgenstern H, Thomas D. Principles of study design in en- vironmental epidemiology. Environ Health Perspect 1993; 101(suppl 4):23–38.

[23] Soto R, Torche A. Spatial inequality, migration, and economic growth in Chile. Cuad Econ 2004;41:401–24.

[24] Axelson O. Aspects of confounding and effect modification in the assessment of occupational cancer risk. J Toxicol Environ Health 1980;6:1127–31.

[25] Henderson SO, Haiman CA, Wilkens LR, et al. Established risk factors account for most of the racial differences in car- diovascular disease mortality. PLoS ONE 2007;2:e377. (DOI: 10.1371/journal.pone.0000377).

[26] Smith AH, Marshall G, Yuan Y, et al. Increased mortality from lung cancer and bronchiectasis in young adults following ex- posure to arsenic in utero and early childhood. Environ Health Perspect 2006;114:1293–6.

[27] Engel RR, Hopenhayn-Rich C, Receveur O, et al. Vascular effects of chronic arsenic exposure: a review. Epidemiol Rev 1994;16:184–209.

[28] Navas-Acien A, Sharrett AR, Silbergeld EK, et al. Arsenic exposure and cardiovascular disease: a systematic review of the epidemiologic evidence. Am J Epidemiol 2005;162: 1037–49.

[29] Hertz-Picciotto I, Arrighi HM, Hu SW. Does arsenic exposure increase the risk for circulatory disease? Am J Epidemiol 2000;151:174–81.

Article 3.4

Increased Childhood Liver Cancer Mortality and Arsenic in Drinking Water in Northern Chile Cancer Mortality in Arsenic-exposed Region II of Chile from 1950 to 2000

Jane Liaw, Guillermo Marshall, Yan Yuan, Catterina Ferreccio, Craig Steinmaus, and Allan H. Smith

University of California, Berkeley, CA, Universidad Católica de Chile and Office of Environmental Health Hazard Assessment, California Environmental Protection Agency, Oakland, California

Abstract. Arsenic in drinking water is an established cause of lung, bladder, and skin cancers in adults and may also cause adult kidney and liver cancers. Some evidence for these effects originated from region II of Chile, which had a period of elevated arsenic levels in drinking water, in particular from 1958 to 1970. This unique exposure scenario provides a rare opportunity to investigate the effects of early-life arsenic exposure on childhood mor- tality; to our knowledge, this is the first study of child- hood cancer mortality and high concentrations of arsenic in drinking water. In this article, we compare cancer mortality rates under the age of 20 in region II during 1950 to 2000 with those of unexposed region V, dividing subjects into those born before, during, or after the peak exposure period. Mortality from the most common childhood cancers, leukemia and brain cancer, was not increased in the exposed population. However, we found that childhood liver cancer mortality occurred at higher rates than expected. For those exposed as young children, liver cancer mortality between ages 0 and 19 was especially high: the relative risk (RR) for males born during this period was 8.9 [95% confidence interval (95% CI), 1.7-45.8; P = 0.009]; for females, the cor- responding RR was 14.1 (95% CI, 1.6-126; P = 0.018); and for males and females pooled, the RR was 10.6 (95% CI, 2.9-39.2; P < 0.001). These findings suggest that exposure to arsenic in drinking water during early childhood may result in an increase in childhood liver cancer mortality. [1]

1 Introduction

Arsenic has been found in drinking water at high levels in many parts of the world, including Bangladesh,India, Argentina, and Chile, and it has been shown to cause numerous health effects, such asskin, bladder, and lung cancer(1-5).Antofagasta,the second largest city in Chile, and neighboring city Mejillones experienced a distinct period of very high arsenic levels in drinking water when their water supply was supplemented in 1958 with water from rivers that contained arsenic at concen- trations near 1,000 μg/L. By comparison, the current WHO recommendation for maximum arsenic concentration in drinking water is 10 μg/L. Before 1958, Anto- fagasta had arsenic concentrations ~90 μg/L, but from 1958 onwards, levels were on average at 870 μg/L until 1971 when an arsenic removal plant was installed. Arsenic levels in the drinking water thus

[1] Liaw J, Marshall G, Yuan Y, Ferreccio C, Steinmaus C, and Smith AH. 2008. Increased Childhood Liver Cancer Mortality and Arsenic in Drinking Water in Northern Chile Cancer Mortality in Arsenic-exposed Region II of Chile from 1950 to 2000. Cancer Epidemiol Biomarkers Prev 2008;17(8):1982–7

dropped suddenly to $\sim 110\mu g/L$ and since then have been further reduced (Fig. 1; ref. 6). Because the area of Antofagasta and Mejillones is extremely dry,there are few individual water sources, and the whole population drinks from the municipal water source. Until very recently, Chile was divided into 13 regions, and Antofagasta and Mejillones are located in region II. Together, these two cities make up more than half of the region II population. All other major cities and towns in region II also had high concentrations of arsenic in region II for varying overlapping periods (2, 6). The large population, distinct period of very high exposure, and well-documented exposure history make region II a highly unique and advantageous area to investigate the health effects of ingested arsenic. The widespread nature and uniformity of the high exposures in this area and the presence of a nearby comparable unexposed area minimize ecologic fallacy and other biases that some researchers commonly associate, sometimes mistakenly, with ecologic studies.

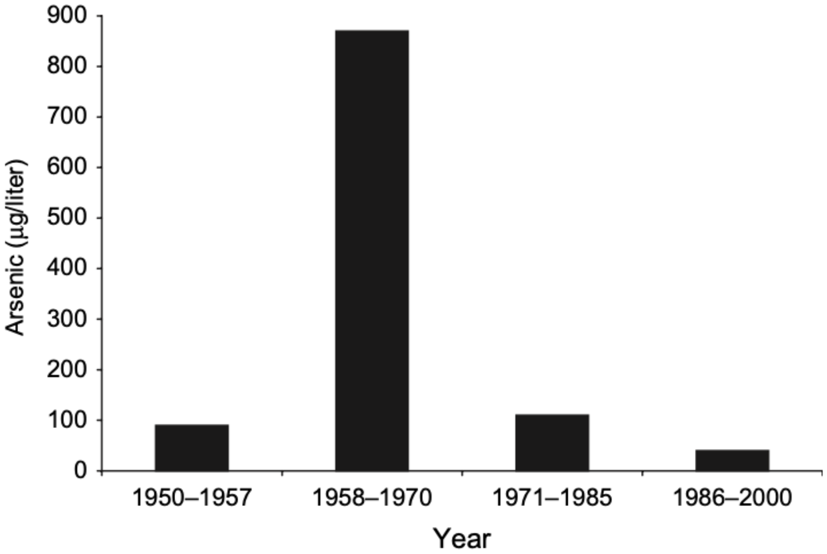
We recently showed that childhood and in utero exposure to arsenic in Antofagasta and Mejillones resulted in markedly increased risks of bronchiectasis and lung cancer in young adults (3). These findings were unexpected, and the magnitude of increased risks from early-life exposure is without precedent. Finding adult cancer resulting from early-life exposure to arsenic led us to also investigate childhood cancer following early-life exposure. The study we report here takes advantage of the unique exposure situation in region II to assess childhood cancers caused by in utero and/or early-life exposure.

2 Materials and Methods

Computerized mortality data were obtained for regions II and V of Chile from the Ministry of Health for the period 1980 to 2000 and from the Chilean National Institute of Statistics (Instituto Nacional de Estadísticas) for the period 1971 to 1979. Due to political unrest in Chile, data from 1976 were unavailable. Computerized mortality data were not available since 1971. For the years 1950 to 1970, death certificates for region II and a referent region were photographed and coded by trained nosologists according to the 9th revision of the International Classification of Diseases (ICD-9). The nosologists were kept blind to the region from which each death certificate originated. Because it was impractical and prohibitively expensive for the study team to code causes of death for all of the death certificates for the entire country of Chile for the years 1950 to 1970, a smaller referent population was chosen. Region V was used as the referent population because of its sociodemographic similarity to Chile as a whole, its population size, and its low levels of arsenic exposure. Region V has had a population about four times the size of the region II population over the course of the study period, from 1950 to 2000. Having a referent population that is significantly larger than that of region II maximizes the statistical precision in the estimation of mortality rate ratios. In preliminary investigations, it was determined that key sociodemographic factors were similar between region V and the rest of Chile. For example, per capita income in region V in 1990 was similar to that of the rest of the country (US$2,053 versus US$2,011). Region V has had low exposures to arsenic in drinking water: in data collected from water companies for the period 1990 to 1994, arsenic water levels in the city of Valparaiso (the largest city in region V) were found to be below the analytic detection limit of $20\mu g/L$,and there is no evidence to suggest that Valparaiso or any other city or town in region V had any past exposures to arsenic in drinking water (7). All of the above information indicates that using region V as a referent population would be a suitable substitute for using all of Chile as a referent population.

Causes of death were coded according to ICD-9 for 1971 to 1998 and according to the 10th revision for the years 1999 to 2000, including leukemia (ICD-9 204-208 and ICD-10 C91-C95),

Fig. 1. Arsenic concentrations in Antofagasta/Mejillones water by year. An arsenic removal plant was installed in 1971.



brain cancer (ICD-9 191 and ICD- 10 C71), and liver cancer (ICD-9 155 and ICD-10 C22). The large majority of death certificates in regionII (89.8%) and region V (94.5%) were certified by physicians. Annual estimates of the population living in regions II and V were obtained for the period 1950 to 2000 from the National Institute of Statistics (Instituto Nacional de Estadísticas), stratified by age and sex. We considered childhood cancer deaths in the range from age 0 to 19. All childhood cancer mortality combined and the two most common childhood cancers, leukemia and brain cancer, were examined first along with all ”other” cancers combined. The ”all other” cancers category displayed unusually elevated relative risks (RR) for girls in region II compared with region V. The individual cancer types in this category (including liver cancer) were therefore examined separately to see if there were increases in some individual cancer sites. We used Poisson regression analysis to calculate RRs between region II and region V mortality rates and the associated 95% confidence intervals (95% CI) for boys and girls separately and combined for the age group 0 to 19 at time of death. Poisson regression analysis was done using the PROC GENMOD procedure provided in Statistical Analysis System software (version 8.2; SAS Institute, Inc.). RRs were calculated for the groups of persons born in 1950 to 1957 (before high exposure), 1958 to 1970 (during high exposure), and 1971 to 1981 (after high exposure), as the observed number of deaths divided by the expected number of deaths, using region V as the referent population.

3 Results

Childhood cancer mortality data for exposed region II and unexposed region V for the years 1950 to 2000 are presented in Table 1 for all childhood cancers combined and for childhood leukemia, brain cancer, and all other cancers. Subjects who were born before 1958 had high exposure as

Table 1. Leukemia, brain cancer, all other childhood cancer, and all childhood (ages 0-19) cancer mortality rates for 1950 to 2000 for children born before high exposure (1950-1957), during high exposure (1958-1970), and after high exposure (1971-1981)

Cause of death	Year of birth	Region II mortality	Region V mortality	Person-years (region II)	Person-years (region V)	RR (95% CI)	*P* Value
All male childhood cancers	1950-1957	35	124	429,640	1,526,848	1.0 (0.7-1.5)	0.99
	1958-1970	41	222	805,712	2,900,314	0.7 (0.5-0.9)	0.02
	1971-1981	43	161	833,134	2,689,807	0.9 (0.6-1.4)	0.51
All female childhood cancers	1950-1957	22	82	433,183	1,528,021	1.0 (0.6-1.5)	0.82
	1958-1970	48	173	801,664	2,872,533	1.0 (0.7-1.4)	0.97
	1971-1981	36	95	813,999	2,609,758	1.2 (0.9-1.6)	0.20
Male brain cancer	1950-1957	2	7	429,640	1,526,848	1.0 (0.2-4.9)	0.98
	1958-1970	1	9	805,712	2,900,314	0.4 (0.1-3.2)	0.38
	1971-1981	3	18	833,134	2,689,807	0.5 (0.1-2.6)	0.44
Female brain cancer	1950-1957	0	8	433,183	1,528,021	0	–
	1958-1970	2	9	801,664	2,872,533	0.8 (0.2-3.7)	0.77
	1971-1981	4	9	813,999	2,609,758	1.4 (1.0-2.1)	0.07
Male leukemia	1950-1957	22	57	429,640	1,526,848	1.4 (0.8-2.2)	0.21
	1958-1970	23	102	805,712	2,900,314	0.8 (0.5-1.3)	0.37
	1971-1981	16	71	833,134	2,689,807	0.7 (0.4-1.5)	0.38
Female leukemia	1950-1957	8	34	433,183	1,528,021	0.8 (0.4-1.8)	0.64
	1958-1970	18	90	801,664	2,872,533	0.7 (0.4-1.2)	0.20
	1971-1981	14	50	813,999	2,609,758	0.9 (0.4-2.1)	0.81
All other male cancers	1950-1957	10	54	429,640	1,526,848	0.7 (0.3-1.3)	0.22
	1958-1970	15	93	805,712	2,900,314	0.6 (0.3-1.0)	0.05
	1971-1981	21	61	833,134	2,689,807	1.1 (0.9-1.4)	0.30
All other female cancers	1950-1957	14	38	433,183	1,528,021	1.3 (0.7-2.4)	0.40
	1958-1970	27	69	801,664	2,872,533	1.4 (0.9-2.2)	0.14
	1971-1981	17	35	813,999	2,609,758	1.6 (1.1-2.3)	0.02

young children, those born between 1958 and 1970 had exposure in utero and as young children, whereas those born after 1970 were never exposed to high levels of arsenic in drinking water.

There was little evidence of any increased cancer risks for all childhood cancers combined nor for the individual sites of brain cancer and leukemia. For leukemia, in males born in 1950 to 1957 just before high exposure, the RR was 1.4 (95% CI, 0.8-2.2), whereas for those born during high exposure and after high exposure the RRs were 0.8 (95%CI, 0.5-1.3) and 0.7(95%CI, 0.4-1.5), respectively. As seen in Table 1, RRs for leukemia in females were all near 1.0 [1950-1957: 0.8 (95% CI, 0.4-1.8); 1958-1970: 0.7 (95% CI,0.4-1.2); 1971-1981: 0.9 (95%CI, 0.4-2.1)]. For all other cancers combined, the RR among females of region II was elevated especially for the time period 1971 to 1981 (RR, 1.6; 9% CI, 1.1-2.3). Because of this unusual finding, we decided to analyze the component cancers for the all other cancer group separately. This led to unexpected findings for liver cancer among both boys and girls.

For both males and females, liver cancer deaths occurred at much higher numbers than expected, especially for those born in 1950 to 1957 just before the high exposure period and who would have been exposed as young children (Table 2; Fig. 2). For males born between 1950 and 1957, the RR was 8.9 (95%CI,1.7-45.8; P = 0.009), whereas for females born between 1950 and 1957, the RR was 14.1 (95% CI, 1.6-126.2; P = 0.018). The pooled RR estimate for boys and girls was 10.6 (95% CI, 2.9-39.3; P ¡ 0.001). For those born between 1958 and 1970 (during the high exposure period), the male liver cancer mortality RR was 1.2 (95% CI, 0.1-11.5), whereas the female liver cancer mortality RR was 1.2 (95% CI, 0.1-11.5). After the period of high exposure, during the years between 1971 and 1981, the male mortality RRw as at 1.1 (95% CI, 0.1-10.4), the female mortality RR was 3.2 (95% CI,0.2-51.3), and the pooled RR was at 1.6 (95%CI, 0.3-8.8; Fig. 2).

All but one of the liver cancer deaths among those born in 1950 to 57 was aged 10 to 19 at the time of death. Thus, there were eight liver cancer deaths in region II aged 10 to 19, and none in region V, which had a population three times larger. The probability of this occurring by chance is less than 1 in 100,000. Childhood liver cancer mortality is normally extremely rare, and finding no comparable deaths in region V is not surprising. From 1972 onwards, we have mortality data for all of Chile, and there were only 13 liver cancer deaths aged 10 to 19 during 1972 to 1982 in the whole of the rest of Chile excluding region II. Using these data to calculate childhood liver cancer rates for the earlier time period in region II, we have estimated the rate ratio in 1950 to 1957 to be 27.3 for boys (9553.7 for girls (95Further information on the eight liver cancer deaths in region II is given in Table 3, including data from death certificates and birth certificates. We could not locate the birth certificates for two cases. The birth certificate of the first case on the list was given as 1949 and not 1950 as calculated from information on the death certificate. We retained this case in some analyses because making corrections based on birth certificates obtained for just five cases could introduce bias in the analyses. However, we also estimated RR with this case excluded: for males born in 1950 to 1957, the RR after the exclusion was 7.11 (95% CI, 1.30-38.8; P = 0.02), and for males and females pooled, RR was 9.44 (95% CI, 2.50-35.6; $P < 0.001$). It is likely that the same 1-year discrepancy in year of birth occurred with others, and making corrections based on birth certificates obtained for five cases would introduce bias in the analyses. Of the eight cases of liver cancer,five were confirmed to have been born in region II, in the cities of Antofagasta (one), Calama (two), Pedro de Valdivia (one), and Tocopilla (one). One case was born in region IV, in the city of Barraza, but resided in Calama in region II for an unknown period before her death. In all region II cities, arsenic levels in drinking water were substantially higher during the years of exposure than the WHO recommended maximum concentration of arsenic in drinking water, 10 μg/L, or levels in the rest of the country. In Pedro de Valdivia and Tocopilla, the arsenic concentration in drinking water from 1950 to 1970 was 250 μg/L, whereas in Calama, the levels were an average of 150 μg/L (8).

Table 2. Male, female, and pooled liver cancer mortality rates from 1950 to 2000 for those born before high exposure (1930-1939, 1940-1949, and 1950-1957), during high exposure (1958-1970), and after high exposure (1971-1981 and 1982-2000) for ages 0 to 19

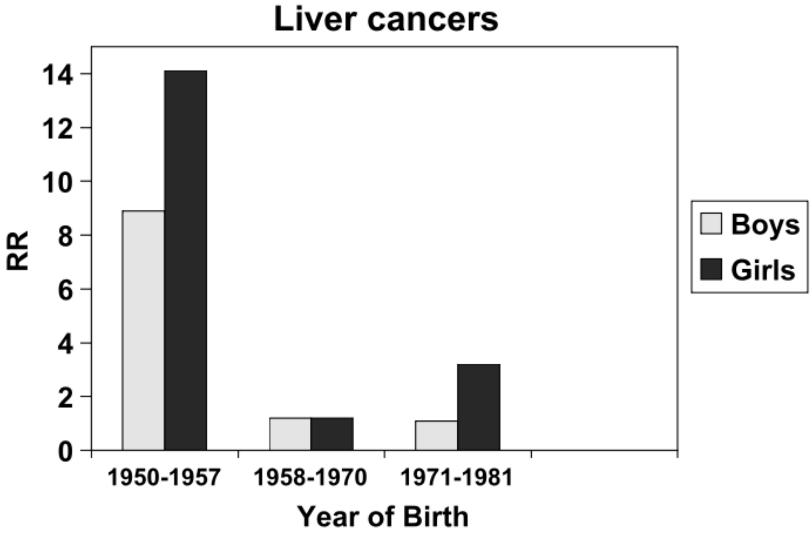
Cause of death	Year of birth	Region II mortality	Region V mortality	Person-years (region II)	Person-years (region V)	RR (95% CI)	*P* Value
Males	1930-1939	2	2	79,145	282,706	3.6 (0.5-25.4)	0.20
	1940-1949	2	4	204,760	752,514	1.8 (0.3-9.8)	0.50
	1950-1957	5	2	429,640	1,526,848	8.9 (1.7-45.8)	0.009
	1958-1970	1	3	805,712	2,900,314	1.2 (0.1-11.5)	0.87
	1971-1981	1	3	833,134	2,689,807	1.1 (0.1-10.4)	0.95
	1982-2000	1	1	826,691	2,489,933	3.0 (0.2-48.2)	0.44
Females	1930-1939	0	1	78,833	292,225	–	–
	1940-1949	1	3	313,431	1,123,930	1.2 (0.1-11.5)	0.88
	1950-1957	4	1	433,183	1,528,021	14.1 (1.6-126.2)	0.018
	1958-1970	1	3	801,664	2,872,533	1.2 (0.1-11.5)	0.88
	1971-1981	1	1	813,999	2,609,758	3.2 (0.2-51.3)	0.41
	1982-2000	0	5	794,560	2,389,023	0	–
Pooled	1930-1939	2	3	157,978	574,931	2.4 (0.4-14.4)	0.33
	1940-1949	3	7	623,139	2,232,402	1.5 (0.4-5.9)	0.53
	1950-1957	9	3	862,823	3,054,869	10.6 (2.9-39.3)	<0.001
	1958-1970	2	6	1,607,376	5,772,847	1.2 (0.2-5.9)	0.83
	1971-1981	1	4	1,647,133	5,299,565	1.6 (0.3-8.8)	0.58
	1982-2000	1	6	1,621,251	4,878,956	0.5 (0.1-4.2)	0.52

Children born in the period 1958 to 1970 would have also experienced some childhood exposure as well as in utero exposure. Overall, there was no evidence of increased risks for the age range 0 to 19 (Table 2). However, focusing on the age range 10 to 19, there was one liver cancer death in region II and one in region V, giving a rate ratio estimate of 3.60 (95% CI, 0.23-57.5). The confidence interval is very wide and no conclusion can be drawn from this risk estimate.

4 Discussion

This is the first study to find evidence that ingestion of arsenic in drinking water in early childhood might increase the risks of childhood liver cancer mortality. Arsenic in drinking water is established to be a major cause of adult cancer in exposed populations (9) and we previously reported marked increases in lung and bladder cancer in the same study population in region II of Chile (10), including increased risks of lung cancer in young adults following early childhood exposure (3). In light of these findings, it is reassuring not to find increased overall risks for childhood cancer mortality. However, among the subjects exposed as young children born in 1950 to 1957 based on death certificate data, we found that liver cancer mortality was markedly increased for both boys and girls aged 0 to 19 [for males, RR was 8.9 (95%CI, 1.7-45.8); for females, the RR was 14.1 (95% CI, 1.6-126.2)]. In the age range of 10 to 19 years, all eight liver cancer cases died in region II, and none in the much larger population of region V, a finding that

Fig. 2. Liver cancer RRs for boys and girls before high exposure period (1950-1957), during high exposure period (1958-1970), and after high exposure period (1971-1981).



had a probability of being due to chance of less than 1 in 100,000. Although the overall number of cases was relatively small, the large exposed and unexposed populations we studied, the high magnitude of the RRs identified, and the low associated P values along with the consistency of findings in both genders all suggest that these findings are not due to chance or bias.

Arsenic exposure has been suggested to increase hepatic cancer in adults, primarily in studies in Taiwan (11). However, we have previously reported that there was no increase in adult liver cancer mortality in region II of Chile (2). Liver cancer is a rare cancer in children and usually comprises only ∼1% of childhood cancers (12). Childhood hepatic cancers comprise two main types, hepatoblastoma, which is the predominant form in children under the age of 5, and hepatocellular carcinoma, which predominates in older children. In a recent United States study, among children 5 years and younger, hepatoblastoma accounted for 91% of cases, whereas hepatocellular carcinoma accounted for 87% of cases among those 15 to 19 years of age (13). In our study, all but one of the children who died from liver cancer was in the age range 10 to 19. A limitation of this study is that we do not have pathology reports, although in two cases hepatocellular carcinoma was the diagnosis given on the death certificate. Because hepatitis B is a major risk factor for childhood hepatocellular carcinoma (13-17), we considered the possibility that the increased risks we identified were due to hepatitis B in region II. However, the pattern of marked increases in liver cancer mortality in both boys and girls coinciding with the very high arsenic exposure period in region II, with subsequent decreases in risks following the end of this high exposure period, suggests that arsenic in water was a more likely explanation. Furthermore, to our knowledge, the relationship between hepatitis B and childhood liver cancer has been in populations with chronic hepatitis B infection rather than occurring in sudden outbreaks.

There is little other information on childhood cancers and exposure to arsenic. We conducted a study of childhood cancer incidence in Nevada and arsenic in drinking water and did not find overall evidence of increased childhood cancer incidence with exposure, and there was only one

Table 3. Liver cancer deaths ages 10 to 19 in region II among those born between 1950 and 1957 in region II of Chile

Gender	Year of birth	Place of birth	Place of death	Year of death	Age of death	Cause of death on death certificate*
M	1950[†]	Tocopilla	Tocopilla	1969	19	Cancer de higado
M	1954	Unknown	Antofagasta	1968	14	Adenocarcinoma hepatocellular
M	1954	Pedro de Valdivia	Antofagasta	1972	18	Cancer hepático
M	1955	Antofagasta	Antofagasta	1967	12	Neo hepatocellular
M	1956	Calama	Calama	1970	13	Cáncer primario de higado
F	1953	Barraza	Antofagasta	1970	18	Cancer hepatico
F	1955	Calama	Calama	1965	10	Cancer de higado, hepatoma
F	1956	Unknown	Region II city unknown	1972	16	Previously ICD coded as liver cancer

Abbreviation: Neo, neoplasm.
* "Higado" is Spanish for liver, "hepático" is Spanish for "of the liver," and "hepatoma" means liver cancer.
[†] Year of birth was derived from death certificate data for the study. When the birth certificate was obtained, it was discovered that this boy was actually born in 1949.
However, identical death certificate data had been used for regions II and V, so the boy was included in data analysis.

case of liver cancer in 20 years (18). However, the highest exposure category involved a range from 35 to 90 μg/L of arsenic in water. We also investigated a remarkable childhood leukemia cluster that occurred in the city of Fallon where water concentrations were ~90 μg/L (19). Eleven cases were diagnosed between 1999 and 2001, resulting in an age-standardized rate ratio in the county of 12.0 (95% CI, 6.0-21.4). However, there was no basis for linking this cluster solely to arsenic in the water because the cluster had occurred only relatively recently, whereas arsenic levels in this area had been at the same concentration for about the last 50 years.

Arsenic exposure was assessed in one population-based case-control study of childhood leukemia based in Quebec, Canada (20). The authors reported slightly increased RRs for childhood acute lymphoblastic leukemia related to postnatal exposure to arsenic (odds ratio, 1.94; 95% CI, 0.64-5.83) but not for prenatal period exposure. These exposures were very low, at an average water concentration of 5μg/L, more than 100 times lower than the exposure of children in region II of Chile in the peak exposure period. Falk et al. (21) reported four female cases of childhood hepatic angiosarcoma, one of whom was exposed to arsenic through multiple sources. Her father worked at a mine,and the child was exposed to arsenic via soil around her house, water, and dust on the father's work clothes. One case of angiosarcoma of the liver has also been reported in region II of Chile diagnosed in 1971 at the age of 22, who also had arsenic-related skin lesions (22). He lived in Pedro de Valdivia like one of the cases we have listed in Table 3. Unfortunately, it is not now possible to obtain medical records to search for pathology reports for the liver cancer deaths in Table 3, all of whom died more than 35 years ago, but as we noted above, two of the death certificates make specific mention of hepatocellular carcinoma and there is no mention of angiosarcoma.

A limitation of this study is the lack of individual data on exposure. However, the evidence suggests that virtually everyone in region II was exposed to much higher concentrations of arsenic in drinking water than in region V. Almost all drinking water in region II is supplied by a few large municipal water sources, and almost the entire region II had high concentrations of arsenic in their drinking water during various periods. The population-weighted average arsenic concentration in drinking water for the entire region was ~580 μg/Lfor around 13 years from 1958 to 1970 (23). Because almost everyone in region II was exposed, ecological fallacy, which can occur if one cannot be sure if those with the outcome were actually exposed, is highly unlikely in this study. In addition, the increases in RRs are too large to result from factors such as migration. In fact, any inmigration of people from other regions of Chile would most probably have diluted the RR effects caused by arsenic in water.

In conclusion, we found marked increases in liver cancer mortality in the age range 10 to 19 among children who had exposure to high concentrations of arsenic in drinking water starting soon after birth. We did not find overall increases in childhood cancer mortality in region II of Chile. The importance of these findings is in part due to region II of Chile having the largest population in the world uniformly exposed to very high concentra- tions of arsenic in drinking water. This gives reassurance that arsenic in water probably does not increase overall childhood cancer mortality. However, the uniqueness of the exposure scenario in region II also means that it will be difficult to identify other large enough populations with high, well-documented arsenic exposure to confirm our findings about childhood liver cancer.

Bibliography

[1] Steinmaus CM, Moore LE, Hopenhayn-Rich C, Biggs ML, Smith AH. Arsenic in drinking water and bladder cancer. Cancer Invest 2000; 18: 174–82.

[2] Smith AH, Goycolea M, Haque R, Biggs ML. Marked increase in bladder and lung cancer mortality in a region of Northern Chile due to arsenic in drinking water. Am J Epidemiol 1998;147:660–9.

[3] Smith AH, Marshall G, Yuan Y, et al. Increased mortality from lung cancer and bronchiectasis in young adults after exposure to arsenic in utero and in early childhood. Environ Health Perspect 2006;114: 1293–6.

[4] Smith AH, Arroyo AP, Mazumder DN, et al. Arsenic-induced skin lesions among Atacameño people in Northern Chile despite good Cancer Epidemiol Biomarkers Prev 2008;17(8). August 2008 Cancer Epidemiology, Biomarkers & Prevention 1987 nutrition and centuries of exposure. Environ Health Perspect 2000; 108: 617–20.

[5] Tondel M, Rahman M, Magnuson A, Chowdhury IA, Faruquee MH, Ahmad SA. The relationship of arsenic levels in drinking water and the prevalence rate of skin lesions in Bangladesh. Environ Health Perspect 1999; 107: 727–9.

[6] Ferreccio C, Gonzalez CA, Milosavjlevic V, Marshall G, Sancha AM, Smith AH. Lung cancer and arsenic concentrations in drinking water in Chile. Epidemiology 2000; 11: 673–9.

[7] Hopenhayn-Rich C, Browning SR, Hertz-Picciotto I, Ferreccio C, Peralta C, Gibb H. Chronic arsenic exposure and risk of infant mortality in two areas of Chile. Environ Health Perspect 2000;108: 667–73.

[8] Rivara MI, Cebria ´n ME, Corey G, Hernandez M, Romieu I. Cancer risk in an arsenic-contaminated area of Chile. Toxicol Ind Health 1997;13:321–38.

[9] International Agency for Research on Cancer. Some drinking-water disinfectants and contaminants, including arsenic. Vol. 84. Lyon (France): WHO; 2004.

[10] Marshall G, Ferreccio C, Yuan Y, et al. Fifty-year study of lung and bladder cancer mortality in Chile related to arsenic in drinking water. J Natl Cancer Inst 2007;99:920–8.

[11] Chiu HF, Ho SC, Wang LY, Wu TN, Yang CY. Does arsenic exposure increase the risk for liver cancer? J Toxicol Environ Health A 2004; 67: 1491–500.

[12] International Incidence of Childhood Cancer, Vol. II. IARC Sci Publ. 1998;144:1–391.

[13] Darbari A, Sabin KM, Shapiro CN, Schwarz KB. Epidemiology of primary hepatic malignancies in U.S. children. Hepatology 2003;38: 560–6.

[14] Chang MH, Shau WY, Chen CJ, et al. Hepatitis B vaccination and hepatocellular carcinoma rates in boys and girls. JAMA 2000;284: 3040–2.

[15] Stiller CA, Pritchard J, Steliarova-Foucher E. Liver cancer in European children: incidence and survival, 1978-1997. Report from the Automated Childhood Cancer Information System project. Eur J Cancer 2006; 42: 2115–23.

[16] Chen JC, Chang ML, Lin JN, etal. Comparison of childhood hepatic malignancies in a hepatitis B hyper-endemic area. World J Gastro- enterol 2005; 11: 5289–94.

[17] Chang MH, Chen TH, Hsu HM, et al. Prevention of hepatocellular carcinoma by universal vaccination against hepatitis B virus: the effect and problems. Clin Cancer Res 2005; 11: 7953–7.

[18] Moore LE, Lu ML, Smith AH. Childhood cancer incidence and arsenic exposure in drinking water in Nevada. Arch Environ Health 2002; 57: 201–6.

[19] Steinmaus CM, Lu ML, Todd RL, Smith AH. Probability estimates for the unique childhood leukemia cluster in Fallon, Nevada, and risks near other U.S. military aviation facilities. Environ Health Perspect 2004; 112: 766–71.

[20] Infante-Rivard C, Olson E, Jacques L, Ayotte P. Drinking water contaminants and childhood leukemia. Epidemiology 2001; 12: 13–9.

[21] Falk HL, Herbert JT, Edmonds L, Heath CW Jr, Thomas LB, Popper H. Review of four cases of childhood hepatic angiosarcoma-elevated environmental arsenic exposure in one case. Cancer 1981; 47: 382–91.

[22] Rennke H, Prat G, Etcheverry R, et al. Hemangioendothelioma maligno del higato y arsenicismo cronico (in Spanish). Rev Med Chil 1971; 99: 664–8.

[23] Yuan Y, Marshall G, Ferreccio C, et al. Acute myocardial infarction mortality in comparison with lung and bladder cancer mortality in arsenic-exposed region II of Chile from 1950 to 2000. Am J Epidemiol 2007; 166: 1381–91.

Article 3.5

Mortality in Young Adults following in Utero and Childhood Exposure to Arsenic in Drinking Water

Allan H. Smith, Guillermo Marshall, Jane Liaw, Yan Yuan, Catterina Ferreccio, and Craig Steinmaus

University of California, Berkeley, CA, Universidad Católica de Chile and Office of Environmental Health Hazard Assessment, California Environmental Protection Agency, Oakland, California

Abstract. Backgroud: Beginning in 1958, the city of Antofagasta in northern Chile was exposed to high arsenic concentrations (870 µg/L) when it switched water sources. The exposure abruptly stopped in 1970 when an arsenic-removal plant commenced operations. A unique exposure scenario like this— with an abrupt start, clear end, and large population (125,000 in 1970), all with essentially the same exposure—is rare in environmental epidemiology. Evidence of increased mortality from lung cancer, bronchiectasis, myocardial infarction, and kidney cancer has been reported among young adults who were in utero or children during the high-exposure period.

Objective: We investigated other causes of mortality in Antofagasta among 30- to 49-year-old adults who were in utero or $\leq$ 18 years of age during the high-exposure period.

Methods: We compared mortality data between Antofagasta and the rest of Chile for people 30–49 years of age during 1989–2000. We estimated expected deaths from mortality rates in all of Chile, excluding Region II where Antofagasta is located, and calculated standardized mortality ratios (SMRs).

Results: We found evidence of increased mortality from bladder cancer [SMR = 18.1; 95% confidence interval (CI): 11.3, 27.4], laryngeal cancer (SMR = 8.1; 95% CI: 3.5, 16.0), liver cancer (SMR = 2.5; 95% CI: 1.6, 3.7), and chronic renal disease (SMR = 2.0; 95% CI: 1.5, 2.8).

Conclusions: Taking together our findings in the present study and previous evidence of increased mortality from other causes of death, we conclude that arsenic in Antofagasta drinking water has resulted in the greatest increases in mortality in adults ¡ 50 years of age ever associated with early-life environmental exposure.[1]

Keywords: arsenic, childhood exposure, Chile, drinking water, environmental exposure, in utero, mortality.

Millions of people worldwide are exposed to arsenic in their drinking water, and arsenic is a well-documented cause of many serious health effects. The International Agency for Research on Cancer (2004) classified arsenic in drinking water as carcinogenic to humans, based on evidence

[1] Smith AH, Marshall G, Liaw J, Yuan Y, Ferreccio C, and Steinmaus C. 2012. Mortality in Young Adults following in Utero and Childhood Exposure to Arsenic in Drinking Water. Environ Health Perspect 120:1527–1531.

that arsenic causes cancers of the skin, lung, and bladder. Chronic arsenic exposure has also been shown to cause noncancer health outcomes in multiple organs, including repro- ductive, cardiovascular, pulmonary, neurologic, and dermal effects (National Research Council 1999, 2001). In the present study we investigated all causes of death following probable in utero and early-life exposure to arsenic.

We examined mortality in an area in northern Chile that has some unique features that make it an ideal location to study long- term outcomes from arsenic exposure. It is the driest inhabited place on earth (McKay et al. 2003). Because it had no private wells, all residents drank water from the only available source: the city water supply. Antofagasta obtained drinking water from rivers that flow from springs in the Andes Mountains. Before 1958, the arsenic level of the city water supply was about 90 μg/L. In 1958, a new city water supply was installed using water from the Toconce and Holajar rivers, which contained 800 and 1,300 μg/L of arsenic, respectively (Smith et al. 1998). With these new sources of water, the average arsenic concentration in the city water supply increased dramatically to 870 μg/L. After a water treatment plant was installed in 1970, the arsenic concentration in the city water supply dropped to 110 μg/L for about 10 years and was reduced further thereafter. The water supply now contains ¡ 10 μg/L of arsenic.

In previous studies of mortality among adults 30–49 years of age, we discovered increased mortality due to lung cancer and bronchiectasis (Smith et al. 2006), kidney cancer (Yuan et al. 2010), and acute myocardial infarction (Yuan et al. 2007) among residents born during or shortly before the high-exposure period. We therefore decided to extend our analysis to assess mortality from all causes of death among adults 30–49 years of age who were born during or before the high-exposure period. Because the age range of 30–49 years is a young age at which to die, we refer to these deaths as deaths in "young adults."

1 Methods

We previously reported the methods used for mortality studies in northern Chile (Smith et al. 2006; Yuan et al. 2010). In brief, we obtained computerized mortality data for 1989–2000 from the Ministry of Health (Santiago, Chile) for all 13 regions of Chile. Deaths were divided into two groups: residents of Antofagasta and neighboring Mejillones (cities located in Region II of Chile that have the same water source), and residents in all other regions of Chile. Antofagasta is very much larger than Mejillones, so here we refer to Antofagasta and Mejillones combined as Antofagasta. We selected 1989 as the first year because it is the first year for which deaths in Antofagasta were reported separately from the rest of Region II. Other parts of Region II also had arsenic in drinking water, although to a lesser extent than Antofagasta. Outside of Region II, the rest of Chile has not been exposed to high levels of arsenic in drinking water. With rare exceptions, arsenic concentrations in water sources outside of Region II have been ¡ 10 μg/L. In a 1984 nationwide survey of 2,000 people, average urine arsenic concentrations were 14 μg/L (Venturino 1991). This is similar to average levels found in the general U.S. population (16.7 μg/L) and indicates very low arsenic concentrations in drinking water (Steinmaus et al. 2009). In addition to mortality data, we obtained survey and census data comparing variables such as smoking, diet, and socioeconomic status from Antofagasta and the rest of Chile in order to evaluate potential confounding from these factors. In the present study, we focused on deaths during 1989–2000 among people 30–49 years of age from Antofagasta or from all of Chile except Region II (referents). People from Antofagasta in this age range would have been in utero, children, or adolescents (up to 18 years of age) during at least part of the high-exposure period of 1958–1970. Two birth cohorts were defined for this investigation: births during 1958–1970 (probable in utero and childhood exposure to those born in Antofagasta), and births during 1940–1957 (probable childhood, but not in utero, exposure to those born in Antofagasta).

The Ministry of Health coded causes of death for 1989–1998 using the International Classification of Diseases, 9th Revision (ICD-9; World Health Organization 1978) and for 1999–2000 using the 10th revision (ICD-10; World Health Organization 1992). For our analysis, we converted all ICD-10 codes into ICD-9 codes. In our initial analysis, we noticed that several causes of death listed under ICD-9 codes 580–589 (genitourinary) exhibited elevated standardized mortality ratios (SMRs). The excess deaths were limited to chronic renal failure (ICD-9 code 585), renal failure unspecified (ICD-9 code 586), chronic glomerulonephritis (ICD-9 code 582), and renal sclerosis (ICD-9 code 587). Because each of these causes of death relates to chronic renal disease and renal failure, we grouped them together as mortality from chronic renal disease.

We obtained annual estimates of the population living in Antofagasta in Region II, as well as for the rest of Chile excluding Region II, for 1989–2000 from the National Institute of Statistics (Instituto Nacional de Estadísticas); estimates were stratified by age and sex. We calculated SMRs for deaths among those 30–49 years of age, using 10-year age groups (30–39 and 40–49 years) for age standardization. We calculated significance and con- fidence intervals (CIs) based on the Poisson distribution (Selvin 1995). In view of the clear direction of the a priori hypotheses for arsenic causing increased risks for both malignant and nonmalignant diseases, here we present one-tailed tests of statistical significance. We tested for effect modification by age group (comparing 30- to 39-year and 40- to 49-year age groups) and effect modification by sex using Poisson regression interaction terms with two-tailed tests. We also tested for effect modification by birth period, comparing mortality for those born in 1940–1949 with those born in 1950–1957, but because there were no significant differences by birth period for any outcome (p ¿ 0.05), we report results for both periods combined. We compared mortality at 30–49 years of age among those born in 1940–1957, who would have experienced at least part of their exposure $\leq$ 18 years of age, with mortality among those born during 1958–1970, most of whom would have been exposed in utero if they were born in Antofagasta.

2 Results

Based on a 1990 random sample survey of Chilean cities that included Antofagasta [Caracterizacion Socio Economica Nacional; (CASEN) 1990], the prevalence of smoking in 1990 was comparable between Antofagasta and the rest of Chile (Table 1). Demographic characteristics for Region II from the 2002 Census (Instituto Nacional de Estadisticas Chile 2002) are similar for Region II and Chile as a whole, including the percentage of the population living in urban areas (98% and 87%, respectively) and the percentage of medically certified death certificates (90% and 86%). Diet and other health risk factors collected in studies of stratified random population census samples conducted in 2003 and 2009 (Gobierno de Chile, Ministerio de Salud 2003, 2010), including obesity, blood cholesterol, glucose, and hypertension, were also similar between the comparison populations. We first estimated SMRs comparing specific causes of death among adults 30–49 years of age born in Antofagasta during 1940–1970 (both sexes combined) to the same age group born in the rest of Chile (data not shown). We observed no increased mortality from infectious and parasitic diseases (ICD-9 codes 001–139; SMR = 1.0; 95% CI: 0.8, 1.3), endocrine and nutritional diseases (ICD-9 codes 240–279; SMR = 1.2; 95% CI: 0.7, 1.8), for diseases of the respiratory system (ICD-9 codes 460–519; SMR = 1.1; 95% CI: 0.9, 1.3), or diseases of the digestive system (ICD-9 codes 520–579; SMR = 0.8; 95% CI: 0.7, 0.9). Mortality was increased for all cancers combined (ICD-9 codes 140–239; SMR = 1.7; 95% CI 1.6, 1.9; $p < 0.001$), deaths from circulatory diseases (ICD-9 codes 390–459; SMR = 1.7; 95% CI: 1.5, 2.0; $p < 0.001$), and diseases of the genitourinary system (ICD-9 codes 580–629; SMR =2.0; 95% CI: 1.5, 2.8; $p < 0.001$).

Table 1. Comparing smoking data, demographic variables, and risk factors for Region II (of which Antofagasta constitutes more than half of the total population) with those for all of Chile.

Variable	Region II	All of Chile
Smoking (%) [a]		
Nonsmokers	78.0	77.5
Moderate smokers (¿ 0 to 1 pack/day)	21.0	21.1
Heavy smokers (¿ 1 pack/day)	1.0	1.2
Men smokers	27.4	26.6
Women smokers	16.6	19.3
Demographic variable (%) [b]		
Urban	98	87
Catholic	72	70
Literate	98	97
Prebasic education	4	4
University/professional education	17	14
Death certificate certified by physician	90	86
Medical risk factor [c]		
Average BMI (kg/cm^2)	27.6	26.8
Obese [BMI ¿ 30 (%)]	19.2	21.9
Morbidly obese [BMI ¿ 40 (%)]	2.8	1.3
Average HDL cholesterol (mg/dL)	34.2	44.6
Average LDL cholesterol (mg/dL)	105.0	115.4
Hypertension [blood pressure ¿ 140/90 (%)]	28.9	33.5
Average total cholesterol (mg/dL)	174.0	186.0
Average serum glucose (mg/dL)	85.8	92.9
Diabetes (%)	3.2	4.2
Regular exercise (%)	13.8	9.2
Dietary risk factor in national survey [d]		
Alcohol consumed per day (g)	41.5	55.6
Fruit/vegetables consumed per day (g)	174.0	186.0
Salt consumption/day (g)	173.8	185.8

Abbreviations: BMI, body mass index; HDL, high-density lipoprotein; LDL,low-density lipoprotein.

[a] Data from CASEN (1990). [b] Data from Instituto Nacional de Estadisticas Chile (2002).

[c] Data from Gobierno de Chile, Ministerio de Salud (2003).

[d] Data from Gobierno de Chile, Ministerio de Salud (2010).

Table 2 presents the SMRs for selected causes of death in Antofagasta for males and females 30–49 years of age during 1989–2000 and the expected numbers estimated from those in the rest of Chile (excluding Region II). The “all other cancers” category comprises all cancers other than those of the bladder, larynx, and liver, as well as lung and kidney cancers, which we previously reported to be associated with early-life arsenic exposure (Smith et al. 2006; Yuan et al. 2010). Mortality for all cancers combined was increased for males born during 1940–1957, most of whom would have experienced at least some exposure before 18 years of age (SMR = 2.1; 95% CI: 1.9, 2.4; $p < 0.001$), and for those born during the high-exposure period (1958–1970), most of whom would have experienced both in utero and childhood exposure (SMR = 2.2; 95% CI: 1.7, 2.8; p < 0.001). Mortality from all cancers combined was also increased among females (SMR = 1.4; 95% CI: 1.2, 1.6 and SMR = 1.4; 95% CI: 1.1, 1.8 for those born in 1940–1957 and 1958–1970, respectively), but to a lesser extent than in males.

Mortality from bladder cancer was greatly increased in males and females (Table 2), particularly among those born during 1958–1970, the high-exposure period with probable exposure in utero (for males, SMR = 65.7; 95% CI: 24.1, 143; for females, SMR = 43.0; 95% CI: 8.9, 126; $p = 0.01$ for interaction between birth periods 1940–1957 and 1958–1970 for males and females combined, adjusted for sex). Increases in liver cancer mortality were also more pronounced for those born during 1958–1970 (for males, SMR = 5.9; 95% CI: 1.9, 13.7; for females, SMR = 4.7; 95% CI: 1.3, 12.0; $p = 0.04$ for interaction by birth period among males and females combined). Mortality from laryngeal cancer was increased among men born in 1940–1957, before the high-exposure period (SMR = 8.9; 95% CI: 3.6, 18.3). Among noncancer causes of death for adults 30–49 years of age, we observed evidence of increased mortality from chronic renal disease, with SMRs in the range of 1.9–2.5 for men and women born during or before the high-exposure period (Table 2). When data for both time periods were combined, we observed no significant differences in SMRs by sex for outcomes evaluated for the first time in the present study, or for outcomes evaluated previously using SMRs for different time periods or separately for men and women [lung cancer, bronchiectasis, and other chronic obstructive pulmonary disease (COPD) (Smith et al. 2006); acute myocardial infarction (Yuan et al. 2007); and kidney can- cer (Yuan et al. 2010)] (Figure 1). The highest combined SMRs were for bladder cancer (SMR = 18.1; 95% CI: 11.3, 27.4) and bronchiectasis (SMR = 18.4; 95% CI: 10.3, 30.4) (Figure 1).

Fig. 1. Summary of SMRs for 30–49-year-old males and females (pooled) who were born in Antofagasta, Chile, combining those born before and during the high-exposure period (Smith et al. 2006; Yuan et al. 2007, 2010)

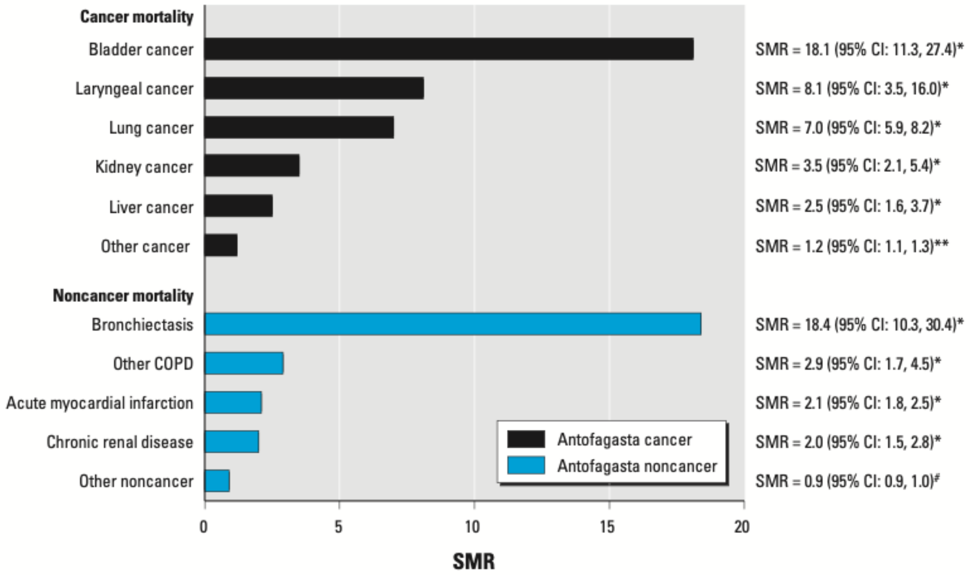


Table 2. Observed and expected deaths and SMRs for males and females 30-49 years of age during 1989-2000 and born in Antofagasta, Chile, in 1940-1957 and 1958-1970 (during the high-exposure period).

		Born 1940-1957				Born 1958-1970				*P*-Value [a]
Cancer	Sex	Observed	Expected	SMR (95% CI)	*P*-Value[a]	Observed	Expected	SMR (95% CI)	P-Value	for interaction[b]
All cancer	Male	226	105.3	2.1 (1.9, 2.4)	<0.001	69	30.8	2.2 (1.7, 2.8)	<0.001	
	Female	219	154.5	1.4 (1.2, 1.6)	<0.001	59	41.3	1.4 (1.1, 1.8)	<0.01	0.83
Bladder cancer	Male	11	0.8	13.7 (6.8, 24.5)	<0.001	6	0.1	65.7 (24.1, 143)	<0.001	
	Female	2	0.3	7.9 (1.0, 28.6)	0.03	3	0.1	43.0 (8.9, 126)	<0.001	0.01
Laryngeal cancer	Male	7	0.8	8.9 (3.6, 18.3)	<0.001	1	0.0	27.4 (0.7, 153)	0.04	
	Female	0	0.1	–	–	0	0.0	–	–	0.53
Liver cancer	Male	10	4.1	2.4 (1.2, 4.4)	0.01	5	0.9	5.9 (1.9, 13.7)	<0.01	
	Female	6	4.1	1.5 (0.5, 3.2)	0.23	4	0.9	4.7 (1.3, 12.0)	0.01	0.04
All other cancers[c]	Male	79	82.9	1.0 (0.8, 1.2)	0.64	41	27.7	1.5 (1.1, 2.0)	0.01	
	Female	174	142.1	1.2 (1.0, 1.4)	<0.01	48	39.0	1.2 (0.9, 1.6)	0.09	0.36
Chronic renal disease[d]	Male	14	7.5	1.9 (1.0, 3.1)	0.02	6	2.7	2.3 (0.8, 4.9)	0.05	
	Female	14	7.1	2.0 (1.1, 3.3)	0.02	6	2.4	2.5 (0.9, 5.4)	0.04	0.71
All other noncancer	Male	310	367.7	0.8 (0.8, 0.9)	0.99	110	128	0.9 (0.7, 1.0)	0.94	
deaths minus injuries[e]	Female	187	178.3	1.0 (0.9, 1.2)	0.27	89	61.8	1.4 (1.2, 1.8)	<0.01	0.33

All data presented for Antofagasta include neighboring Mejillones, which had the same water sources.

[a] One-sided *P*-value. [b] Two-sided *P*-value; test for interaction between birth periods 1940-1957 and 1958-1970 adjusted for sex.

[c] "All other cancers" comprises all cancers except those of the bladder, larynx, liver, lung, and kidney. [d] ICD-9 codes 580-589.

[e] "All other noncancer deaths minus injuries" comprises all noncancer deaths except injuries, acute myocardial infarction, bronchiectasis, and other COPD; the last three of these
diseases were previously shown to be associated with early-life arsenic exposure (Smith et al. 2006; Yuan et al. 2007).

3 Discussion

We found increases in mortality from several different causes of death in the 30- to 49-year-old study population. We previously reported increases in lung cancer and bronchiectasis in Chile following early-life arsenic exposure (Smith et al. 2006), as well as increases in bronchiectasis in India after adult exposure (Guha Mazumder et al. 2005). We have also reported increases in mortality from kidney cancer (Yuan et al. 2010) and myocardial infarction (Yuan et al. 2007) after early-life exposure to arsenic in Antofagasta. Bladder cancer SMRs from the present analysis for adults (30–49 years of age) born during the high-exposure period (1958–1970; males: SMR = 65.7; 95% CI: 24.1, 143, females: SMR = 43.0; 95% CI: 8.9, 126) are 5–10 times higher than those we reported previously for all ages combined, regardless of age at exposure (males, 6.0; females, 8.2) (Smith et al. 1998).

The association between early-life exposure to arsenic and mortality from laryngeal cancer suggests that this cancer might be related to arsenic exposure, although the number of laryngeal cancer deaths was very small and the association was evident only among men.

The findings concerning mortality from chronic renal disease were unexpected. In one study in Taiwan, Chiu and Yang (2005) reported a relationship between arsenic and mortality from renal diseases, but the authors did not specifically investigate early-life exposure. In the study area in Taiwan, residents had consumed arsenic-contaminated well water with a median concentration of 780 μg/L and had renal disease mortality rates that were 50% higher than those in unexposed populations. It is plausible that renal disease might relate to arsenic in water because arsenic is excreted through the kidney and is also a probable cause of kidney cancer (National Research Council 2001; Yuan et al. 2010).

Adult liver cancer has been linked to arsenic exposure, primarily in studies in Taiwan (Chiu et al. 2004). In a previous study in Region II of Chile, we found little evidence of increased mortality from liver cancer in adults (Smith et al. 1998), but we did not evaluate early-life exposure. However, in a subsequent analysis of childhood cancers (Liaw et al. 2008), liver cancer mortality in children 10–19 years of age was increased in Region II compared with Region V [for males, relative risk (RR) = 8.9; 95% CI: 1.7, 45.8; p = 0.009; for females, RR = 14.1; 95% CI: 1.6, 126; p = 0.018]. The data we present here are, to our knowledge, the first to link early-life arsenic exposure to liver cancer mortality in adults 30–49 years of age.

A major strength of this study is the large size of the exposed population: There were > 125,000 residents in Antofagasta and Mejillones in 1970 who were exposed to water arsenic concentrations of 870 μg/L, including approximately 60,000 children with early-life exposure during the high-exposure period. The largest cohort study on arsenic conducted in Taiwan involved only 698 subjects $\geq$ 40 years of age exposed to arsenic concentrations > 300 μg/L (Chiou et al. 2001). Recently published cohort studies in Bangladesh involved 10,431 subjects $\geq$ 15 years of age exposed to arsenic at 300 μg/L, the largest study (Sohel et al. 2009), and 2,889 subjects $\geq$ 18 years of age exposed to 150μµg/L, in the second largest study (Argos et al. 2010). In addition, the populations studied in Taiwan, India, and Bangladesh received their water from a large number of small-town or domestic wells with wide variations in arsenic concentrations, even between closely located wells (Guha Mazumder et al. 1998; Van Geen et al. 2002), making it extremely difficult to accurately estimate early-life exposure decades earlier.

One potential weakness of the present study is that it is ecological, comparing Antofagasta with the rest of Chile. However, this study does not have the usual problems associated with what is sometimes termed the “ecologic fallacy” (Morgenstern 1995) because the one source of drinking water in Antofagasta had a known concentration of arsenic; therefore, we can be confident that virtually everyone who lived in Antofagasta during the high-exposure period was indeed exposed. Migration into and out of the study area could also introduce bias, but people migrating from Antofagasta to elsewhere in Chile would constitute a very small proportion of the total Chilean population. Any resulting bias would tend to reduce relative mortality estimates for Antofagasta. Migration into Antofagasta also would tend to bias estimates toward the null because these residents would be misclassified as exposed even though they did not reside in Antofagasta during the high-exposure period. In addition, migration within Chile is relatively uncommon: From 1965 to 2000, annual internal migration among regions of Chile was only 0.6% compared with 1.2% in Argentina, 3.1% in the United Kingdom, and 6.6% in United States (Soto and Torche 2004). Another potential weakness of the study is the use of death certificate data. Based on Chile's Census 2002 data (Instituto Nacional de Estadisticas Chile 2002), most death certificates (86%) in the country are signed by physicians; in Region II, where Antofagasta is located, the percentage is similar (90%). In addition, Chile has a national health care system that services the whole country (Reichard 1996). Therefore, it is unlikely that differences in diagnostic practices between Antofagasta and the rest of Chile would produce spurious differences in mortality rates. The importance of ecological studies in causal inference concerning arsenic in drinking water was recognized by the International Agency for Research on Cancer (2004):

> For most other known human carcinogens, the major source of causal evidence arise from case–control and cohort studies, with little, if any, evidence from ecological studies. In contrast, for arsenic in drinking-water, ecological studies provide important information on causal inference, because of large exposure contrasts and limited population migration.

There are two reasons why confounding is not a concern, the first reason of which involves the magnitude of the mortality rate ratios identified. Consider, for example, the SMRs for acute myocardial infarction mortality of 2.3 and 2.7 among Antofagasta men for births during 1940–1957 and 1958–1970, respectively. These values are comparable to the acute myocardial infarction mortality rate ratio of 2.11 for current cigarette smokers who smoked ¿ 20 pack-years compared with never-smokers from a large cohort study of about 140,000 men in the United States (Henderson et al. 2007). These similarities suggest that confounding by smoking would explain our SMRs only if all men in Antofagasta smoked but no men in the rest of Chile smoked. Similarly, confounding by smoking would not explain the estimated SMRs for lung cancer, laryngeal cancer, and bladder cancer.

The second reason confounding is not a major concern is that there is no evidence of major differences in risk factors (other than arsenic) between Antofagasta and the rest of Chile. For example, a survey conducted in 1990 indicated that the prevalence of smoking in Region II (27.4% in men and 16.6% in women) was similar to Chile as a whole (26.6% in men and 19.3% in women) (Marshall et al. 2007). It is also extremely unlikely for other confounding factors, including diet or exercise, to produce the magnitude of the SMRs we report here. As shown in Table 1, other cardiovascular mortality risk factors, including body mass index, obesity, cholesterol, and hypertension, were not substantially different between Region II and all of Chile in 2003 (Gobierno de Chile, Ministerio de Salud 2003). Having given consideration to all potential sources of bias, we conclude that our study provides strong epidemiological evidence of increased mortality risks from several causes in young adults exposed to arsenic in early life. Inorganic arsenic and its metabolites readily pass through the placenta, exposing the fetus to concentrations similar to those of the mother (Concha et al. 1998). Animal experiments have shown that arsenic is a transplacental carcinogen in mice and causes tumors in offspring (Tokar et al. 2011; Waalkes et al. 2007). Vahter (2008) has shown that arsenic acts epigenetically and interferes with DNA methylation. Ren et al. (2010) reported that arsenic exposure may alter DNA methylation, globally affecting the expression of multiple genes; this may explain why exposure to arsenic is associated with multiple disease outcomes in different organs.

We know of no childhood environmental exposure that results in comparable increases in adult mortality rates (Figure 1). Yorifuji et al. (2010) reported that mortality from pancreatic cancer and leukemia were increased in young adults after arsenic exposure from contaminated milk powder in Japan; these exposures were very high and resulted in acute poisoning effects. In a study of 60,182 people, Vineis et al. (2005) reported elevated lung cancer risks from childhood passive smoking, but the relative risk from "daily, many hours" of passive smoking exposure was 3.63 (95% CI: 1.19, 11.11) and there were only four cases of lung cancer in this group. Other studies of passive childhood smoking have not found increased risks (Boffetta et al. 2000). A nonenvironmental exposure—radiation treatment of childhood cancer—causes major increases in later mortality from other cancers (excluding recurrence of the treated cancer) as well as from noncancer outcomes. A recent report from the Childhood Cancer Survivor Study showed that mortality from other cancers was increased [relative risk (RR) = 2.9; 95% CI: 2.1, 4.2] and that mortality from cardiac causes (RR = 3.3; 95% CI: 2.0, 5.5) and "other" causes (RR = 2.0; 95% CI: 1.3, 3.1) were also increased (Mertens et al. 2008). Increased cancer mortality has also been demonstrated in atomic-bomb survivors exposed in utero or as young children (Preston et al. 2008). Excluding these fairly rare and specific high-dose radiation exposure scenarios, our findings suggest that early-life exposure to arsenic in drinking water results in greater increases in mortality in adults ¡ 50 years of age than those attributable to any other early-life environmental exposure.

4 Conclusions

To our knowledge, this is the first investigation of all causes of death in young adults follow- ing early-life exposure to arsenic in drinking water. In those exposed to water arsenic concentrations of approximately 870 μg/L, we identified pronounced increases in mortality in young adults 30–49 years of age from cancers of the bladder, larynx, and liver and from renal diseases associated with chronic renal failure. Taken together with the increased mortality from other causes (Smith et al. 2006; Yuan et al. 2007, 2010), the magnitude and extent of the increased mortality we have identified are without precedent for any early-life environmental exposure. Our findings need to be confirmed in other populations, but they add strong support for efforts to reduce population exposure to arsenic in drinking water, particularly during pregnancy and childhood.

Bibliography

[1] Argos M, Kalra T, Rathouz PJ, Chen Y, Pierce B, Parvez F, et al. 2010. Arsenic exposure from drinking water, and all-cause and chronic-disease mortalities in Bangladesh (HEALS): a prospective cohort study. *Lancet* 376(9737):252-258.

[2] Boffetta P, Tredaniel J, Greco A. 2000. Risk of childhood cancer and adult lung cancer after childhood exposure to passive smoke: a meta-analysis. *Environ Health Perspect* 108:73-82.

[3] CASEN. 1990. Illa Encuestra CASEN (Caracterizacion Socio Economica Nacional). Santiago, Chile: Ministerio de Planificacion y Cooperacion Nacional Republica de Chile.

[4] Chiou HY, Chiou ST, Hsu YH, Chou YL, Tseng CH, Wei ML, et al. 2001. Incidence of transitional cell carcinoma and arsenic in drinking water: a follow-up study of 8,102 residents in an arseniasis-endemic area in northeastern Taiwan. *Am J Epidemiol* 153(5):411-418.

[5] Chiu HF, Ho SC, Wang LY, Wu TN, Yang CY. 2004. Does arsenic exposure increase the risk for liver cancer? *J Toxicol Environ Health A* 67(19):1491-1500.

[6] Chiu HF, Yang CY. 2005. Decreasing trend in renal disease mortality after cessation from arsenic exposure in a previous arseniasis-endemic area in southwestern Taiwan. *J Toxicol Environ Health A* 68(5):319-327.

[7] Concha G, Vogler G, Lezcano D, Nermell B, Vahter M. 1998. Exposure to inorganic arsenic metabolites during early human development. *Toxicol Sci* 44(2):185-190.

[8] Gobierno de Chile, Ministerio de Salud. 2003. Resultados 1 Encuesta de Salud. Available: `http://epi.minsal.cl/epi/html/invest/ENS/InformeFinalENS.pdf` [accessed 24 August 2012].

[9] Gobierno de Chile, Ministerio de Salud. 2010. Encuesta Nacional de Salud Chile 2009-2010. Available: `http://www.minsal.gob.cl/portal/url/item/bcb03d7bc28b64dfe040010165012d23.pdf` [accessed 19 September 2012].

[10] Guha Mazumder DN, Chakraborty D, et al. 1998. Arsenic levels in drinking water and the prevalence of skin lesions in West Bengal, India. *Int J Epidemiol* 27(5):871-877.

[11] Guha Mazumder DN, Steinmaus C, Bhattacharya P, von Ehrenstein OS, Ghosh N, Gotway M, et al. 2005. Bronchiectasis in persons with skin lesions resulting from arsenic in drinking water. *Epidemiology* 16(6):760-765.

[12] Henderson SO, Haiman CA, Wilkens LR, Kolonel LN, Wan P, Pike MC. 2007. Established risk factors account for most of the racial differences in cardiovascular disease mortality. *PLoS One* 2(4):e377; doi: 10.1371/journal.pone.0000377 [Online 18 April 2007].

[13] Instituto Nacional de Estadisticas Chile. 2002. National Census 2002. Available: `http://www.ine.cl/cd2002/index.php` [accessed 23 August 2012].

[14] International Agency for Research on Cancer. 2004. Some Drinking-water Disinfectants and Contaminants, including Arsenic. *IARC Monogr Eval Carcinog Risks Hum* 84:1-477.

[15] Liaw J, Marshall G, Yuan Y, Ferreccio C, Steinmaus C, Smith AH. 2008. Increased childhood liver cancer mortality and arsenic in drinking water in northern Chile. *Cancer Epidemiol Biomarkers Prev* 17(8):1982-1987.

[16] Marshall G, Ferreccio C, Yuan Y, Bates M, Steinmaus C, Selvin S, et al. 2007. Fifty-year study of lung and bladder cancer mortality in Chile related to arsenic in drinking water. *J Natl Cancer Inst* 99(12):920-928.

[17] McKay CP, Friedmann EI, Gomez-Silva B, Caceres-Villanueva L, Andersen DT, Landheim R. 2003. Temperature and moisture conditions for life in the extreme arid region of the Atacama desert: four years of observations including the El Niño of 1997-1998. *Astrobiology* 3(2):393-406.

[18] Mertens AC, Liu Q, Neglia JP, Wasilewski K, Leisenring W, Armstrong GT, et al. 2008. Cause-specific late mortality among 5-year survivors of childhood cancer: the Childhood Cancer Survivor Study. *J Natl Cancer Inst* 100(19):1368-1379.

[19] Morgenstern H. 1995. Ecologic studies in epidemiology: concepts, principles, and methods. *Annu Rev Public Health* 16:61-81.

[20] National Research Council. 1999. Arsenic in Drinking Water. Washington, DC: National Academy Press.

[21] National Research Council. 2001. Arsenic in Drinking Water: 2001 Update. Washington, DC: National Academy Press.

[22] Preston DL, Cullings H, Suyama A, Funamoto S, Nishi N, Soda M, et al. 2008. Solid cancer incidence in atomic bomb survivors exposed in utero or as young children. *J Natl Cancer Inst* 100(6):428-436.

[23] Reichard S. 1996. Ideology drives health care reforms in Chile. *J Public Health Policy* 17(1):80-98.

[24] Ren X, McHale CM, Skibola CF, Smith AH, Smith MT, Zhang L. 2011. An emerging role for epigenetic dysregulation in arsenic toxicity and carcinogenesis. *Environ Health Perspect* 119:11-19.

[25] Selvin S. 1995. Poisson regression analysis. In: Practical Biostatistical Methods. Belmont, CA: Duxbury Press, 455-496.

[26] Smith AH, Goycolea M, Haque R, Biggs ML. 1998. Marked increase in bladder and lung cancer mortality in a region of northern Chile due to arsenic in drinking water. *Am J Epidemiol* 147(7):660-669.

[27] Smith AH, Marshall G, Yuan Y, Ferreccio C, Liaw J, von Ehrenstein O, et al. 2006. Increased mortality from lung cancer and bronchiectasis in young adults after exposure to arsenic in utero and in early childhood. *Environ Health Perspect* 114:1293-1296.

[28] Sohel N, Persson LA, Rahman M, Streatfield PK, Yunus M, Ekström EC, et al. 2009. Arsenic in drinking water and adult mortality: a population-based cohort study in rural Bangladesh. *Epidemiology* 20(6):824-830.

[29] Soto R, Torche A. 2004. Spatial inequality, migration, and economic growth in Chile. *Cuadernos de Economia* 41:401-424.

[30] Steinmaus C, Yuan Y, Liaw J, Smith AH. 2009. Low-level population exposure to inorganic arsenic in the United States and diabetes mellitus: a reanalysis. *Epidemiology* 20(6):807-815.

[31] Tokar EJ, Diwan BA, Ward JM, Delker DA, Waalkes MP. 2011. Carcinogenic effects of "whole-life" exposure to inorganic arsenic in CD1 mice. *Toxicol Sci* 119(1):73-83.

[32] Vahter M. 2008. Health effects of early life exposure to arsenic. *Basic Clin Pharmacol Toxicol* 102(2):204-211.

[33] Van Geen A, Ahsan H, Horneman AH, Dhar RK, Zheng Y, Hussain I, et al. 2002. Promotion of well-switching to mitigate the current arsenic crisis in Bangladesh. *Bull World Health Organ* 80(9):732-737.

[34] Venturino P. 1991. Determinacion de concentracion de arsenico urinario en diferentes regiones de Chile. In: Primera Jornada Sobre Arsenicismo Laboral y Ambiental (Ministerio de Salud. Servicio de Salud Antofagasta, Departamento de programas sobre el ambiente yACdS, ed). Antofagasta, Chile: Republica de Chile, 49-52.

[35] Vineis P, Airoldi L, Veglia P, Olgiati L, Pastorelli R, Autrup H, et al. 2005. Environmental tobacco smoke and risk of respiratory cancer and chronic obstructive pulmonary disease in former smokers and never smokers in the EPIC prospective study. *BMJ* 330(7486):277; doi: 10.1136/bmj.38327.648472.82 [Online 3 February 2005].

[36] Waalkes MP, Liu J, Diwan BA. 2007. Transplacental arsenic carcinogenesis in mice. *Toxicol Appl Pharmacol* 222(3):271-280.

[37] World Health Organization. 1978. Manual of the International Statistical Classification of Diseases, Injuries, and Causes of Death: Based on the Recommendations of the Ninth Revision Conference, 1975, and Adopted by the Twenty-ninth World Health Assembly, Vol 1. Geneva: World Health Organization.

[38] World Health Organization. 1992. International Statistical Classification of Diseases and Related Health Problems, 10th Revision. Geneva: World Health Organization.

[39] Yorifuji T, Tsuda T, Grandjean P. 2010. Unusual cancer excess after neonatal arsenic exposure from contaminated milk powder [Letter]. *J Natl Cancer Inst* 102(5):360-361.

[40] Yuan Y, Marshall G, Ferreccio C, Steinmaus C, Selvin S, Liaw J, et al. 2007. Acute myocardial infarction mortality in comparison with lung and bladder cancer mortality in arsenic-exposed Region II of Chile from 1950 to 2000. *Am J Epidemiol* 166(12):1381-1391.

[41] Yuan Y, Marshall G, Ferreccio C, Steinmaus C, Selvin S, Liaw J, et al. 2010. Kidney cancer mortality: fifty-year latency patterns related to arsenic exposure. *Epidemiology* 21(1):103-108.

Chapter 4

Statistical Methods for Quality of Care

Article 4.1

Prospective Prediction in the Presence of Missing Data

Guillermo Marshall, Bradley Warner, Samantha MaWhinney, and Karl E. Hammermeister

Pontificia Universidad Católica de Chile,
University of Colorado Health Sciences Center, and
Department of Veterans Affairs Medical Center

Abstract. A variety of methods and algorithms are available for estimating parameter in the class of generalized linear model in presence of missing values. However, there is little information on how this already built model can be used for prediction in new observations with missing data in the covariates. Dropping the observations with missing values is a widespread practice with serious statistical and non-statistical implications. One solution is to fit separate regression models, or sub-models, to each pattern of missing covariates. In practice, for any iterative regression method, this approach is computationally intensive. We propose a simple methodology to predict outcomes for individual with incomplete information based on the estimated coefficients and its covariance from the already built model. This method does not require to revisit the original dataset used to built the original model and works by generating a first-order approximation of any sub-model coefficient estimates. This is achieved by using the SWEEP operator on an augmented covariance matrix obtained from the original model. We refer to this approach as the *one step sweep* (OSS) method.

The methodology is demonstrated using data from the Department of Veterans Affairs Continuous Improvement in Cardiac Surgery Program (CICSP). These data contain 30 day mortality, the outcome of interest, and risk information for over 14,000 patients who underwent coronary artery bypass grafting (CABG) surgery over a four year period. Using complete data from the first 3.5 years of this study period, a logistic regression model was built. This model was then used to predict mortality for patients undergoing CABG in the most recent 6 months. In order to evaluate the performance of the OSS method we randomly generated observations with missing covariates in the 6 month prediction database. We use this simulation to demonstrate that the computationally efficient OSS substantially reduces the error in risk-adjusted mortality created when cases with incomplete information are eliminated. Lastly, we derive the relationship between the OSS method and data imputation.[1]

Keywords: missing values, prediction, generalized linear models, logistic regression, risk adjusted mortality rate, measuring quality of health care

1 Introduction

Generalized linear models provides a rich class of tools for modeling a wide range of outcome variables, including continuous and categorical responses, and their association with a set of

[1] Marshall G, Warner B, Samantha MaWhinney, and Hammermeister KE. 2002. Prospective Prediction in the Presence of Missing Data. Statistics in Medicine 21 (4), 561-570

covariates. It is often of interest in clinical and epidemiological studies to investigate the relationship between a binary response representing the presence or absence of a disease and a set of patient risk factors.

Classical statistical method require to have complete information in both the outcome and the covariate for all the subject included in the study in order to estimate the parameters. However, in practice these studies often have missing observation in some of the covariates for some of the patients. To deal with this problem there is a variety of methods and algorithm for estimating the parameters in presence of missing values. These methods are either model based (1; 2) or data based (3; 4).

Once the parameters of a model are estimates with one of these methods, the model can be used to predict the outcome in future coming patients. It is well known that these predictive models supplement clinical knowledge and are an important tool in medical decision making (5; 6; 7; 8; 9). Additionally, when data is aggregated by clinician or hospital provides a measure to assess and improve quality of care.

However, because the patient data are abstracted from the patient's medical record they are frequently incomplete and therefore the model can't be used and a predictive outcome for that patient will be also missing. When data is aggregated by clinician or hospital the standard procedure is to drop these patients from the analysis.

In this paper we introduce a computationally efficient method for prospective prediction for observation with missing covariates. We refer to this approach as the one step sweep (OSS) method. The OSS method can be viewed as a data based imputation procedure (see Section **??** for details). The method is implemented in one program that calculates predicted outcomes for all patients in the prediction set, including those with missing covariates. Other techniques require a separate step for data imputation before the prediction step. All that is required to implement this method are the estimates of the regression coefficients for the full predictive model based on a complete data set and the respective estimate of the covariance matrix.

Although this method can be used in other member of the class of generalized linear model and many predictive model self-contained, we will consider the special case of a logistic regression model.

Initially, we build the full model based on observations (patients) with no missing data and retain the estimates of the coefficients and their associated covariance matrix. The expected outcome for an observation with missing covariates is obtained by estimating the sub-model that includes only the observed predictors. The sub-model is estimated using a non-iterative one step approximation based on the full model's coefficients and covariance matrix. As an alternative a separate regression could be used to generate the sub-models, but at a substantial computational cost. Also, for databases with numerous observations, the OSS method does not require the storage of a large design matrix. For our motivating example, the regression requires a 11408×14 design matrix compared to a 15×15 matrix for the OSS method. In this paper, we will focus on the implementation of this method specifically to logistic regression models, although it can be extended to other generalized linear models (10).

The basis for the OSS methodology is the work by Lawless and Singhal (11) developed for performing stepwise regression on non-normal models. Their focus is on model building and variable selection relating to the choice of an appropriate number of predictor variables from a large set of potential candidates. Our method predicts outcomes when some observations have missing covariates.

In the Department of Veterans Affairs (VA) Continuous Improvement in Cardiac Surgery Program (CICSP), the expected 30 day mortality (probability of death) for a patient undergoing cardiac surgery is calculated based on a logistic regression model (12). Every six months for each VA hospital performing cardiac surgery, a quality of care measure is generated using a ratio of the observed deaths (O) to the cumulative expected mortality (E). These O/E ratios along with other outcome data, such as volume and resource utilization, are returned to the cardiac surgical centers for their internal quality improvement review. In addition, these results are reviewed by the VA Cardiac Surgery Consultants Committee, a national quality oversight body. Until recently patients were dropped from the calculation of an O/E ratio if they had missing information on one or more of the predictor variables used in the logistic model (9).

In the VA CICSP Program, nurses were hired specifically to collect data and complete records are available on roughly 98% of patients. However, other national databases contain significantly more missing data. Even though the records for this VA database are nearly complete, it is still important not to drop patients from the analysis. Because death is a relatively rare outcome in cardiac surgery, dropping a death will significantly impact the O/E ratio. Conversely, dropping a patient with a high expected mortality who survives can also affect the O/E ratio, although not to the same extent as a patient who dies.

In Section 2 we begin by developing the methodology using the logistic regression model. Then in Section 3 we show the relationship between the OSS method and the imputation of missing covariates. Lastly, in Section 4 computer simulations are used to evaluate and compare methods.

2 Methodology

In logistic regression the probability of an adverse outcome, π, is related to a vector of covariates $\mathbf{x}$ using the logistic transformation

$$\log \frac{\pi}{1-\pi} = \boldsymbol{\beta}'\mathbf{x}, \tag{1}$$

where $\boldsymbol{\beta}$ is a vector of regression coefficients. The regression coefficients, $\hat{\boldsymbol{\beta}}$, are estimated using maximum likelihood theory and have an asymptotic multivariate normal distribution

$$\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \boldsymbol{V}),$$

where

$$\boldsymbol{V} = \phi(\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X})^{-1}$$

is the covariance matrix. In this expression, ϕ is the dispersion parameter (which equals 1 for logistic regression with no over-dispersion), $\boldsymbol{X}$ is the $n \times p$ design matrix associated with the regression coefficients, and $\boldsymbol{W}$ is the $n \times n$ weight matrix with diagonal elements $\pi(1-\pi)$.

For an observation with missing data on those covariates associated with the regression coefficients $\boldsymbol{\beta}_2$, we can formulate the sub-model

$$\log \frac{\pi}{1-\pi} = \boldsymbol{\beta}_1'\mathbf{x}_1 + 0\mathbf{x}_2, \tag{2}$$

and obtain maximum likelihood estimates for $\boldsymbol{\beta}_1$. If we apply this idea to n new observations, we must fit and save the information from all possible models corresponding to the different missing data patterns. This task is computationally intensive and difficult to implement.

An alternative is to partition the vector of estimates as

$$\begin{pmatrix} \hat{\boldsymbol{\beta}}_1 \\ \hat{\boldsymbol{\beta}}_2 \end{pmatrix} \sim N\left(\begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{pmatrix}, \begin{bmatrix} \boldsymbol{V}_{11} & \boldsymbol{V}_{12} \\ \boldsymbol{V}_{21} & \boldsymbol{V}_{22} \end{bmatrix} \right).$$

and recognize that the conditional distribution of $\hat{\boldsymbol{\beta}}_1$ given $\hat{\boldsymbol{\beta}}_2 = 0$ as

$$\hat{\boldsymbol{\beta}}_1 | \hat{\boldsymbol{\beta}}_2 = 0 \sim N\left(\boldsymbol{\beta}_1 - \boldsymbol{V}_{12}\boldsymbol{V}_{22}^{-1}\boldsymbol{\beta}_2, \boldsymbol{V}_{11} - \boldsymbol{V}_{12}\boldsymbol{V}_{22}^{-1}\boldsymbol{V}_{21} \right). \tag{3}$$

We propose an estimate of $\boldsymbol{\beta}_1$ for the sub-model (2) based on the conditional distribution (3) as

$$\tilde{\boldsymbol{\beta}}_1 = \hat{\boldsymbol{\beta}}_1 - \hat{\boldsymbol{V}}_{12}\hat{\boldsymbol{V}}_{22}^{-1}\hat{\boldsymbol{\beta}}_2 \tag{4}$$

where $\hat{\boldsymbol{\beta}}$ is the maximum likelihood estimate of $\boldsymbol{\beta}$ and $\hat{\boldsymbol{V}}$ is the estimated covariance matrix of the coefficient estimates obtained from the full model (1).

The value of $\tilde{\boldsymbol{\beta}}_1$ can be considered as a one step approximation of the maximum likelihood estimate of $\boldsymbol{\beta}_1$ for sub-model (2). To accomplish this task for all new observations with missing data we introduce a computationally efficient algorithm that uses the SWEEP operator (13).

The first step of this algorithm is to construct a symmetric $(p+1) \times (p+1)$ matrix $\boldsymbol{A}$ with the estimates of $\boldsymbol{\beta}$ and $\boldsymbol{V}$ from the full model (1) as

$$\boldsymbol{A} = \begin{bmatrix} \hat{\boldsymbol{V}} & \hat{\boldsymbol{\beta}} \\ \hat{\boldsymbol{\beta}}' & 0 \end{bmatrix}.$$

Now, given a new observation with complete information for $\boldsymbol{x}_1 = (x_1, x_2, \ldots, x_q)'$ and missing data for $\boldsymbol{x}_2 = (x_{q+1}, x_{q+2}, \ldots, x_p)'$, the estimate of $\boldsymbol{\beta}_1$ proposed in equation (4) can be obtained by applying the sweep operator to the columns $q+1, q+2, \ldots, p$ of the $\boldsymbol{A}$ matrix. The result of this operation is the matrix

$$\boldsymbol{B} = \mathcal{S}_{q+1} \cdots \mathcal{S}_p \boldsymbol{A} = \begin{bmatrix} \hat{\boldsymbol{V}}_{11} - \hat{\boldsymbol{V}}_{12}\hat{\boldsymbol{V}}_{22}^{-1}\hat{\boldsymbol{V}}_{21} & \hat{\boldsymbol{V}}_{12}\hat{\boldsymbol{V}}_{22}^{-1} & \hat{\boldsymbol{\beta}}_1 - \hat{\boldsymbol{V}}_{12}\hat{\boldsymbol{V}}_{22}^{-1}\hat{\boldsymbol{\beta}}_2 \\ \hat{\boldsymbol{V}}_{22}^{-1}\hat{\boldsymbol{V}}_{21} & \hat{\boldsymbol{V}}_{22}^{-1} & \hat{\boldsymbol{V}}_{22}^{-1}\hat{\boldsymbol{\beta}}_2 \\ \hat{\boldsymbol{\beta}}_1' - \hat{\boldsymbol{\beta}}_2'\hat{\boldsymbol{V}}_{22}^{-1}\hat{\boldsymbol{V}}_{21} & \hat{\boldsymbol{\beta}}_2'\hat{\boldsymbol{V}}_{22}^{-1} & -\hat{\boldsymbol{\beta}}_2'\hat{\boldsymbol{V}}_{22}^{-1}\hat{\boldsymbol{\beta}}_2 \end{bmatrix} \tag{5}$$

where $\mathcal{S}_k\boldsymbol{A}$ denote a sweep operation over column k on $\boldsymbol{A}$. Note that in the resulting matrix $\boldsymbol{B}$ we obtain in the upper right corner of the matrix, $\tilde{\boldsymbol{\beta}}_1$, the approximate maximum likelihood estimate of the sub-model coefficients (2). One of the advantages of this algorithm is that the sweep operator is reversible and invariant to the order in which the operations are applied. Applying the sweep operator to a column of the matrix $\boldsymbol{A}$ for a variable that has been removed adds that variable back into the regression, that is $\boldsymbol{A} = \mathcal{S}_k\mathcal{S}_k\boldsymbol{A}$ and $\mathcal{S}_2\mathcal{S}_1\boldsymbol{A} = \mathcal{S}_1\mathcal{S}_2\boldsymbol{A}$.

It is important to point out that sub-models are only approximation to averages of the correct full models. For example if a binary predictor X_2 is missing and if the full model is correctetly specified and includes other predictors $\boldsymbol{X}_1$, the correct predicted value would be

$$\hat{\pi} = Pr\{Y = 1|\boldsymbol{X}_1, X_2 = 0\}Pr\{X_2 = 0|\boldsymbol{X}_1\} + Pr\{Y = 1|\boldsymbol{X}_1, X_2 = 1\}Pr\{X_2 = 1|\boldsymbol{X}_1\}$$

where Y is the binary response variable.

3 Relationship with Imputation

The OSS method introduced in the previous section can be viewed as a data imputation procedure (2). We refer to data imputation as a general term for methods that fill in missing covariates so that all observations can be used in the analysis. At first look, it may appear that the OSS method ignores missing data by building sub-models that exclude missing covariates; however, it can also be considered a data imputation method because it is filling in missing values.

Consider again an observation that has complete data in $\boldsymbol{x}_1 = (x_1, x_2, \ldots, x_q)'$ and missing data for $\boldsymbol{x}_2 = (x_{q+1}, x_{q+2}, \ldots, x_p)'$. Let $\boldsymbol{C}$ be the $(p+1) \times (p+1)$ symmetric matrix obtained after applying the SWEEP operator over the first p columns of the $\boldsymbol{A}$ as

$$\boldsymbol{C} = \mathcal{S}_1 \cdots \mathcal{S}_p \boldsymbol{A} = \begin{bmatrix} -\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X} & \boldsymbol{X}'\boldsymbol{W}\boldsymbol{X}\hat{\boldsymbol{\beta}} \\ \hat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X} & -\hat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X}\hat{\boldsymbol{\beta}} \end{bmatrix}.$$

which when partitioned as before is

$$\boldsymbol{C} = \begin{bmatrix} -\boldsymbol{X}_1'\boldsymbol{W}\boldsymbol{X}_1 & -\boldsymbol{X}_1'\boldsymbol{W}\boldsymbol{X}_2 & \boldsymbol{X}_1'\boldsymbol{W}\boldsymbol{X}\hat{\boldsymbol{\beta}} \\ -\boldsymbol{X}_2'\boldsymbol{W}\boldsymbol{X}_1 & -\boldsymbol{X}_2'\boldsymbol{W}\boldsymbol{X}_2 & \boldsymbol{X}_2'\boldsymbol{W}\boldsymbol{X}\ \hat{\boldsymbol{\beta}} \\ \hat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X}_1 & \hat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X}_2 & -\hat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X}\hat{\boldsymbol{\beta}} \end{bmatrix}.$$

If we sweep on the first q columns of $\boldsymbol{C}$, we obtain $\boldsymbol{B}$ defined in expression (5) due to the reversibility property of the SWEEP operator. In terms of the $\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X}$ notation, we have

$$\boldsymbol{B} = \mathcal{S}_1 \cdots \mathcal{S}_q \boldsymbol{C} = \begin{bmatrix} \left(\boldsymbol{X}_1'\boldsymbol{W}\boldsymbol{X}_1\right)^{-1} & -\hat{\boldsymbol{\alpha}} & \hat{\boldsymbol{\beta}}_1 + \hat{\boldsymbol{\alpha}}\hat{\boldsymbol{\beta}}_2 \\ -\hat{\boldsymbol{\alpha}}' & -\boldsymbol{S} & \boldsymbol{S}\hat{\boldsymbol{\beta}}_2 \\ \hat{\boldsymbol{\beta}}_1' + \hat{\boldsymbol{\beta}}_2'\hat{\boldsymbol{\alpha}}' & \hat{\boldsymbol{\beta}}_2'\boldsymbol{S} & -\hat{\boldsymbol{\beta}}_2'\boldsymbol{S}\hat{\boldsymbol{\beta}}_2 \end{bmatrix},$$

where

$$\hat{\boldsymbol{\alpha}} = \left(\boldsymbol{X}_1'\boldsymbol{W}\boldsymbol{X}_1\right)^{-1} \boldsymbol{X}_1'\boldsymbol{W}\boldsymbol{X}_2$$

and

$$\boldsymbol{S} = \left(\boldsymbol{X}_2 - \boldsymbol{X}_1\hat{\boldsymbol{\alpha}}\right)' \boldsymbol{W} \left(\boldsymbol{X}_2 - \boldsymbol{X}_1\hat{\boldsymbol{\alpha}}\right).$$

The expression for $\hat{\boldsymbol{\alpha}}$ is the weighted least squares estimator of the vector of regression coefficients $\boldsymbol{\alpha}$ in the underlying multivariate linear model

$$\boldsymbol{X}_2 = \boldsymbol{X}_1\boldsymbol{\alpha} + \epsilon \tag{6}$$

weighted by the binomial variance. S is the $(p-q) \times (p-q)$ residual sum of squares matrix. Thus Equation (4) can be written as $\tilde{\boldsymbol{\beta}}_1 = \hat{\boldsymbol{\beta}}_1 + \hat{\boldsymbol{\alpha}}\hat{\boldsymbol{\beta}}_2$.

From the full model an estimate of π with complete data $\boldsymbol{x}' = (\boldsymbol{x}_1', \boldsymbol{x}_2')$ is

$$\hat{\pi} = \left(1 + \exp\left\{-\hat{\boldsymbol{\beta}}_1\boldsymbol{x}_1 - \hat{\boldsymbol{\beta}}_2\boldsymbol{x}_2\right\}\right)^{-1}. \tag{7}$$

However, when data is missing in $\boldsymbol{x}_2$ the OSS estimate is

$$\begin{aligned} \tilde{\pi} &= \left(1 + \exp\left\{-\tilde{\boldsymbol{\beta}}_1\boldsymbol{x}_1\right\}\right)^{-1} \\ &= \left(1 + \exp\left\{-\hat{\boldsymbol{\beta}}_1\boldsymbol{x}_1 - \hat{\boldsymbol{\beta}}_2\hat{\boldsymbol{\alpha}}\boldsymbol{x}_1\right\}\right)^{-1} \\ &= \left(1 + \exp\left\{-\hat{\boldsymbol{\beta}}_1\boldsymbol{x}_1 - \hat{\boldsymbol{\beta}}_2\tilde{\boldsymbol{x}}_2\right\}\right)^{-1}. \end{aligned}$$

where $\tilde{\boldsymbol{x}}_2 = \hat{\boldsymbol{\alpha}}\boldsymbol{x}_1$ is the best linear predictor of $\boldsymbol{x}_2$ based on the complete information $\boldsymbol{x}_1$ using the multivariate model (6).

4 Simulation

The OSS method was tested on a large data set from the VA CICSP. This program collects clinical data from the 40 VA centers performing cardiac surgery. The data were divided into a model building set and a prediction set. The model building set had 11,408 observations, 12 independent covariates and a binary outcome variable, 30 day mortality. The 2,706 observations of the prediction set were used as new data to simulate the calculation of O/E ratios. This process simulates the actual six-month calculation of O/E ratios used at the VA.

The prediction data set had complete information for all covariates; however, to simulate missing data, a program was written to randomly generate missing predictors in this set. The number of covariates in an individual patient selected to have missing values was randomly generated to be between 1 and 4 inclusively. The random generating process was established such that the probability distribution for missing 1,2,3 and 4 covariates was 0.55, 0.25, 0.15 and 0.05, respectively. The process of selecting which of the twelve covariates were missing was also random. From the original prediction data set of 2,706 observations, three new data sets were generated that had 5%, 20% and 50% of the observations with missing covariates.

Four methods were compared for each simulation run on the prediction data set. The comparison of these four methods was done with respect to the results obtained from the logistic model with complete observations. In the first method we generated all logistic regression sub-models for the different missing covariate patterns. That is, if an observation has missing data in a covariate, the logistic regression sub-model that does not include that covariate is built and then used to predict the outcome. The second method was the OSS method described in this paper. The third method imputed the missing covariates using Buck's method of data imputation (3) and then used the logistic regression prediction model as if all data had been observed. The fourth method deleted an observation if any of its covariates were missing, a common practice for missing data and the method previously employed by the VA.

Although, we are comparing four methods of dealing with missing data, for practical purposes the only two methods we believe that could realistically be implemented are the OSS method and the case deletion method. The main reason is that the other methods require large amounts of information to be stored and this significantly limits their portability.

Because it is the measure used by the VA to describe hospital performance, the focus of comparison was the O/E ratio computed by each of these methods. For each of the hospitals in the prediction set, a set of O/E ratios were calculated for each method. The performance of a method was evaluated by comparing its estimated O/E ratios with those from the complete data model. The disagreement between the O/E ratios calculated for the 40 hospitals using one of the four method described above and the O/E ratios calculated using the complete data model was measured using sum of squared differences

$$SSE_{jk} = \sum_{i=1}^{40} \left[\left(\frac{O}{E}\right)_{ijk} - \left(\frac{O}{E}\right)_{i0k} \right]^2 \tag{8}$$

with $j = 0, 1, 2, 3$ corresponding to complete data ($j = 0$), 5% ($j = 1$), 20% ($j = 2$), and 50% ($j = 3$) missing data and $k = 1, 2, 3, 4$ corresponding to all sub-models, OSS, Buck's imputation, and case deletion methods respectively.

Table 1 shows the values obtained for the SSE for each simulation level. As expected, the values of SSE increase with the percentage of missing data. The results for the all sub-models method, the OSS method, and Buck's imputation method show similar and very good performance with respect to the results obtained with complete data, even in the case of a high percentage of missing data. The case deletion method shows significantly more disagreement with respect to results obtained with complete data and this situation becomes dramatically worse when the percent of missing data increases.

Table 1. The SSE for the four alternative methods with three different percentage of missing data

	% Missing Data		
Method	5%	25%	50%
All sub models	0.003	0.013	0.141
OSS	0.003	0.017	0.156
Buck's Imputation	0.003	0.011	0.143
Case Deletion	0.234	4.650	21.95

Figure 1 contains graphs of the O/E ratios from the complete data model versus those obtained from the OSS and case deletion methods. The 45° line represents perfect agreement between a method and the complete data model. These graphs represent the scenarios where 5%, 20%, and 50% of the observations had missing data, respectively. This figure demonstrates that the case deletion method dramatically increases the spread around the 45° line. The figure also shows that the OSS method is robust to O/E ratio error in presence of missing data. Similar results were found for the other two methods.

A significant difference between O/E ratios based on the complete data and those obtained when observations with missing covariates are deleted occurs when a patient who died has missing covariates. Consider the simulation results for a hospital where 5% of the observations had a missing covariate. The O/E ratios for the OSS method and the case deletions method are shown in Table 2. This is clearly a case where one of the adverse outcomes had a missing covariate. In the model where the observation was dropped, the O/E ratio is roughly half of what it was with complete information. This grossly changed the perceived performance of the hospital. The OSS method gave a value similar to the complete data model.

Table 2. Comparison of O/E ratios for the case deletion method and the OSS method. This data represents one hospital with 65 patients and 2 death, where 5% of the observations had missing values for some of the covariates. One of the observations with missing data was a patient who died.

Method	**O/E Ratio**
Complete Data	0.6868
OSS	0.6728
Case Deletion	0.3693

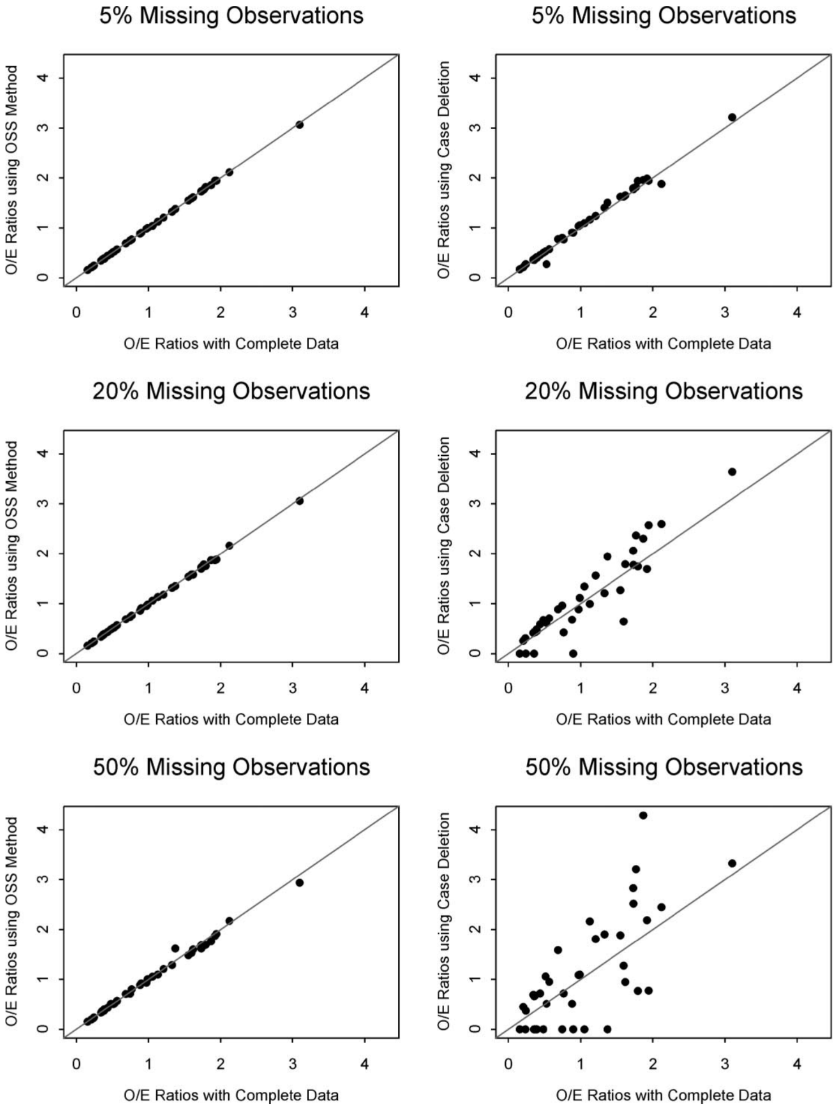


Fig. 1. Comparison of O/E ratios computed using the OSS and case deletion methods for 40 VA hospitals against the complete data model when 5%, 20%, and 50% of the observations have missing values.

Although, the SSE of the first three methods are extremely close, the OSS method is much faster and requires less storage space than generating all logistic regression sub-models. Using the same algorithm and the data where 50% of the observations had missing values, the OSS method obtained estimates in 26 seconds while the complete logistic method took 15 hours on a SUN Sparc Station 20. In addition, the OSS method can be implemented during the data step that calculates expected values and does not require an additional data imputation step.

5 Discussion

We have demonstrated that the OSS method is a computationally efficient approach for obtaining expected values for generalized linear models when observations are missing covariates. As compared with the present use of case deletion, the OSS method significantly reduces the error in the O/E ratios calculated for the VA CICSP database. The OSS results were comparable to those obtained by building complete logistic regression sub-models and data imputation. However, the OSS approach is computationally faster and requires less storage space than building separate sub-models for each permutation of available covariates. Although, the efficiency of the algorithm can be improved by sorting the observations in the new data set based on the missing data patterns, the SWEEP operator is so simple to apply, that the cost of sorting the observations would be similar to the posterior gain obtained in the OSS algorithm.

Although the Continuous Improvement in Cardiac Surgery Program in the Department of Veterans Affairs has nurse data coordinators to collect high quality and complete data, they have decided to implement the OSS method in their semi-annual analysis. Most other hospital systems that are using risk-adjusted mortality rates based on data extracted from medical charts where the percent of missing data is significantly more could also use this new method to help reduce the error in estimating the quality of care of each hospital.

Acknowledgements

This research was supported by the Health Services Research and Development Service, Veterans Health Administration, USA Federal Government and by the Sciences and Technology Foundation of Chile, grant Fondecyt 1980862

Bibliography

[1] Dempster, A., Laird, N., and Rubin, D. Maximum likelihood from incomplete data via the em algorithm, *Journal of the Royal Statistical Society, Series B*, **39**, 1–38, (1977).

[2] Little, R. and Rubin, D. *Statistical Analysis with Missing Values*, John Wiley & Sons, New York, 1987.

[3] Buck, S. A method of estimation of missing values in multivariate data suitable for use with an electronic computer, *Journal of the Royal Statistical Society, Series B*, **22**, 302–306, (1960).

[4] Harrell Jr., F. *Regression Modeling Strategies with Applications to Linear Models, Logistic Regression, and Survival Analysis*, Springer-Verlag, New York, 2001.

[5] Grover, F., Hammermeister, K., and Daley, J. Using outcomes data to improve clinical practice: What we have learned, *The Annals of Thoracic Surgery*, **58**, 1809–1811, (1994).

[6] Hammermeister, K., Shroyer, L., and Dauber, I., Provider perceptions in using outcomes data to improve clinical practice, *The Annals of Thoracic Surgery*, **58**, 1877–1880, (1994).

[7] Iezzoni, L. *Risk Adjustment for Measuring Health Care Outcomes*, Health Administration Press, Ann Arbor, 1994.

[8] Spiegelhalter, D.J. Probabilistic prediction in patient management, *Statistics in Medicine*, **5**, 421–433, (1986).

[9] Marshall, G., Grover, F., Henderson, W., and Hammermeister, K., Assessment of predictive models for binary outcomes: An empirical approach using operative death from cardiac surgery, *Statistics in Medicine*, **13**, 1501–1511, (1994).

[10] McCullagh, P. and Nelder, J. *Generalized Linear Models.* Chapman and Hall, London, 1989.

[11] Lawless, J. and Singhal, K. Efficient screening of nonnormal regression models. *Biometrics*, **34**, 318–327, (1978).

[12] Hammermeister, K., Johnson, R., Marshall, G., and Grover, F. Continuous assessment and improvement in quality of care: A model from the department of veterans affairs cardiac surgery *Annals of Surgery*, **219**, 281–290,(1994).

[13] Goodnight, J. A tutorial on the SWEEP operator, *The American Statistician*, **33**, 149–158, (1979).

Article 4.2

Time Series Monitors of Outcomes

Guillermo Marshall, A. Laurie W. Shroyer, Fred L. Grover, and Karl E. Hammermeister

Pontificia Universidad Católica de Chile,
University of Colorado Health Sciences Center, and
Department of Veterans Affairs Medical Center

Abstract. Despite the popularity of risk-adjusted outcomes as quality of health care indicators, their instability over time and inability to provide reliable comparisons of small volume providers have raised questions about the feasibility and credibility of using these measures. In this paper we describe a new analytical strategy to address these problems by examining risk-adjusted mortality over time, "Time Series Monitors of Outcome" (TSMO), and its application to cardiac surgery performed throughout the Department of Veterans Affairs (VA) between April 1987 and September 1992.

Expected operative mortality for 24,029 patients undergoing coronary artery bypass surgery at all 43 centers performing this procedure was estimated using a logistic regression model to adjust for patient-specific risk factors. The ratio of observed-to-expected (O/E) operative mortality was calculated for each hospital for each of the 11 six-month periods. Poisson regression models were used to identify high and low outlier hospitals based on significant deviation from the 5.5 year overall mean and/or the individual hospital's trend of O/E ratios over time.

This method identified four high and one low outlier hospitals based on significant deviations from the overall mean and three upward and seven downward trending outlier hospitals based on significant deviations in trend over time. A significant downward trend in O/E ratios of 4% per year was also observed for all CABG procedures performed throughout the VA during the last 5.5 year period.

TSMO should help reduce misclassification of outliers due to random variation in outcomes as well as provide more reliable comparative information from which to evaluate provider performance. [1]

Keywords: risk of patient mortality; clinical risk factors; patient outcomes; longitudinal data analysis ; time trends; Poisson regression; and coronary artery bypass graft surgery

1 Introduction

Defining and measuring quality of hospital care is the first step in attempting to improve it. A problem continues to exist in finding a common definition of quality as well as quality measurement tools that are both valid and reliable. Historically, hospital quality of care has primarily been evaluated by clinical case review of adverse events. The effectiveness of the peer review process in identifying quality of care problems and solutions is questionable (1; 2; 3). Furthermore,

[1] G Marshall, ALW Shroyer, FL Grover, KE Hammermeister (1998) Time series monitors of outcomes: a new dimension for measuring quality of care. Medical Care 36 (3), 348-356

clinical case review is time-consuming, expensive, and may have unintended negative effects by virtue of its focus on adverse outcomes (4; 5). Inter-facility comparisons, based on case reviews, are difficult at best. (6).

Raw mortality has been used to monitor quality of patient care within hospital systems for nearly a century (7). Raw mortality does not account for differences among hospitals due to severity of illness and comorbidity of the patient population served. These differences in patient mix across hospitals can introduce significant bias when raw mortality data is used as a measure for hospital quality of care (8).

During the last decade a renewed focus has been placed on the use of mortality rates for comparison of hospital outcomes (9; 10; 11; 12; 13; 14; 15; 16; 17; 18; 19; 20; 21; 22; 23). More recently, there has been an increasing interest in using patient risk-adjusted mortality indicators as a measure of quality of care for hospitals and surgeons (24). The Department of Veteran's Affairs (VA) and the state of New York have taken the initiative in using hospital and surgeon risk-adjusted cardiac surgery mortality rates with the goal of improving the quality of patient care (25; 26). More recently, the states of New York and Pennsylvania have released risk-adjusted cardiac surgery mortality rates by provider for the purpose of providing information on quality of care to the public (27; 28). These initiatives have captured the attention of the media and have created an increasing controversy among the medical community (29). Other states are starting similar initiatives that will likely increase the national attention and public awareness of the quality of our health care system (30; 31)

Hospitals and/or surgeons with observed operative mortality significantly higher or lower than expected are identified as either "high" or "low" outliers, respectively (32) Within the VA cardiac surgery program, an internal report which identifies high and low outlier cardiac surgery programs is prepared semi-annually for use by the VA Cardiac Surgery Consultants Committee.

This six-month period analysis (also called cross-sectional analysis throughout this paper) of comparative hospital data has been very useful to the Cardiac Surgery Consultant Committee. However, marked variability in risk-adjusted mortality rates has been observed for individual hospitals. In fact, we have observed individual hospitals with a high O/E ratio in one period and a normal O/E ratio in the next. These inconsistencies have resulted in both discussion and confusion among the Cardiac Surgery Consultant Committee members in further refining the VA cardiac surgery standards used for performance review and have raised questions regarding the credibility of these measures.

We believe that an important reason for such inconsistencies is the natural variability of these measures and the lack of statistical power to detect moderate and large deviation in risk-adjusted mortality rates. The minimum volume per hospital needed (with a statistical power of 80% and significance level of 5%) to detect a risk-adjusted mortality twice the VA average is approximately 185 cases in a six-month period, a volume achieved by only one hospital on one occasion during the study period of 5.5 years. Other studies have also documented substantial annual variation in hospital cardiac surgical procedure mortality rates (33; 34; 35) Furthermore, since mortality rates for CABG surgery appear to be generally decreasing over time, this overall trend must be considered in any model used to identify "outlier" facilities.

TSMO was developed in direct response to these issues as an analytic tool to increase the statistical power and increase the confidence in the detection of outlier facilities. We believe that by capturing the systematic variation in risk-adjusted mortality rates of individual hospitals over time we can partially address some of these concerns. As a new dimension to the existing measures of hospital performance, an intra-hospital longitudinal analysis of trends in individual hospital

O/E ratios over time, entitled ”Time Series Monitors of Outcome” (TSMO), was developed and is the subject of this report.

2 Methods

The Continuous Improvement in Cardiac Surgery Study

In 1972 the Department of Medicine and Surgery of the DVA Central Office appointed the DVA Cardiac Surgery Consultants Committee to address quality of care and cost effectiveness issues for all cardiac surgery performed within this medical care system. Composed of six cardiac surgeons and three cardiologists from both within and outside the DVA, this committee reviewed procedure-specific operative volume and operative mortality for each of the 43 medical centers performing cardiac surgery. If operative mortality for a six-month period at a medical center exceeded pre-specified limits, the director of that cardiac surgery program was required to prepare summaries of all operative deaths, as well as provide relevant medical records for review by the Committee, which then recommended changes to improve the processes and structures of care. If operative mortality continued to be excessive, a site visit by a team of Committee members could result, and/or the program be closed (36). Recognizing the limitation of using unadjusted operative mortality as an indicator of quality of care, the Cardiac Surgery Consultants Committee encouraged the development of the Cardiac Surgery Risk Assessment Program (now known as the Continuous Improvement in Cardiac Surgery Study) (32; 37) to adjust postoperative outcome on each patient undergoing cardiac surgery for preoperative risk factors. This program started data collection in April 1987.

Patient Population

During the five and one-half years (April 1987 through September 1992) of operation of the Department of Veterans Affairs Cardiac Surgery Risk Assessment Program, data forms were received on 30,330 patients undergoing cardiac surgery from 43 VA medical centers. These initial efforts in developing TSMO were confined to the 24,029 patients undergoing coronary artery bypass grafting. Separate risk models for operative mortality have been developed for the other surgical procedures.

Risk Variables

A total of 54 patient preoperative risk factors were identified initially as potential predictors of operative death based on previous research studies, clinical judgment, and likelihood of complete data collection. After three years of experience, the number of variables collected was reduced to only those factors showing a significant clinical or statistical association with the primary outcome. The 26 risk factors meeting this criterion during the study period were: gender, age, chronic obstructive pulmonary disease, cardiomegaly by chest x-ray, pulmonary rales, current smoker, serum creatinine, rest angina, time since previous percutaneous transluminal coronary angioplasty, time since previous myocardial infarction, prior heart surgery, peripheral vascular disease, New York Heart Association functional class, current diuretic use, current digoxin use, intravenous nitroglycerin with 48 hours preceding surgery, preoperative use of the intra-aortic balloon, left ventricular end-diastolic pressure, aortic systolic pressure, pulmonary artery systolic pressure, mean pulmonary artery wedge pressure, percent left main coronary artery stenosis, number of major coronaries with stenosis(es) * 50%, left ventricular contraction grade or ejection fraction, physician's preoperative estimate of operative mortality, and surgical priority (elective, urgent and emergent).25, 32 These data elements are being collected for every patient undergoing coronary artery bypass graft surgery within the VA hospital system. For patients operated before September 1990, the majority of data was collected by a member of the cardiology/cardiac surgery team without additional funding support. Starting in the Spring of 1991, nurse data coordinators were funded to collect risk and outcome data on approximately 50,000 non cardiac surgeries and 7,000 cardiac surgeries annually. These data coordinators were also asked to retrospectively complete missing cardiac surgery data back to October 1990.

3 Statistical Methods

The first step in developing this new methodological tool was the construction of a predictive model of operative mortality for all 11 six-month periods using all 24,029 patient records. A multiple logistic regression model was developed using stepwise procedures using all 25 preoperative clinical risk factors as predictive variables and 30 days operative mortality as the binary outcome. The computation of this logistic regression was done using the procedure LOGISTIC of the SAS 38 statistical program. The risk factors found to have an independent association with operative mortality as a result of the stepwise procedure were used in the multiple logistic regression model to predict expected operative mortality for each of the 24,029 individual patients.

Aggregated expected mortality by hospital and six-month period was used to calculate the ratio of observed-to-expected operative mortality. High and low outlier hospitals based on one single period of time were detected using exact 95% confidence intervals using a procedure that is based on the binomial distribution. 39 These values, obtained from a cross-sectional analysis, were used to check the consistency of risk-adjusted mortality rates for one time period in comparison with the new longitudinal analysis. The TSMO methodology considers the hospital as the unit of analysis and models variations of O/E ratios over time using a Poisson regression model. 40 The model establishes that the observed number of deaths in each hospital and for each period can be represented by a log-linear model

$$\log \mu_{ij} = \log E_{ij} + \alpha + \beta_i (t_j - \bar{t}) \qquad (1)$$

where μ_{ij} is the expected number of deaths unadjusted for patient-specific risk factors for the ith hospital in the jth period of time, E_{ij} is the expected number of deaths adjusted by risk factors for the same hospital and the same period of time, (α_i, β_i) are the intercept and slope of the assumed linear trend of the O/E ratio overtime in log-scale, and t_j and $\bar{t}$ are the values and the mean respectively of the 11 time periods $t_j = 1, 2, \ldots, 11$. The values of are calculated by adding the predicted operative mortality of all patients undergoing cardiac surgery in the ith hospital and in the jth time period using our multiple logistic regression model previously developed. The values of $\exp\{\alpha_i\}$ and $\exp\{\beta_i\}$ represent the mean O/E ratio over time and the mean change between two consecutive six-month periods for the ith hospital respectively. A value of $\alpha_i = 0$ indicates a mean O/E ratio over time of one, and a value of $\beta_i = 0$ indicates no systematic variation over time in O/E ratios. Hospitals can be identified as high, normal, or low outliers based on either the mean O/E ratio or the trend in O/E ratio over time according to the result of individual 95% confidence intervals for the parameters α_i and β_i, respectively. The Wald test or likelihood-based methods can be used to create confidence intervals for the two indicators α_i and β_i, of this model (40) The classification of hospitals according to low, normal, and high mean

O/E ratios or downward, stable, and upward O/E ratio trends based on the estimates of α_i and β_i produce a total of nine possible categories as shown in Table 1.

Table 1. Classification of 43 VA medical center as high, normal, or low outliers by significant deviation from the overall mean and trend in risk-adjusted mortality over time.

	Trend			
Mean	Downward	Stable	Upward	Total
Low	1	0	0	1
Normal	6	30	2	38
High	0	3	1	4
Total	7	33	3	43

A common slope model can be introduced to represent the overall trend in O/E ratio for the entire VA during the last five and one-half years. Departures from this linear trend were tested by introducing a Poisson additive model (41). A Poisson additive model is an extension of a Poisson regression model that allow the covariable, in this case the time, to be represented by a smooth curve instead of a straight line (Figure 1). Auto-correlation of the errors was also added into the model using the method described by Liang and Zeger (42). The method proposed by these authors extend traditional Poisson regression model allowing the observations of an individual hospital to be correlated from one period to another. The likelihood ratio test (LRT) was used to compare these alternative models with the model assuming linear trend. The GENMOD procedure of SAS were used to fit the Poisson regression models (43).

4 Results

The Predictive Model for Risk Adjustment

Stepwise logistic regression models have been demonstrated by a previous study (44) to be an efficient approach to develop a predictive model for risk adjustment. In this study, the predictive model was validated by splitting the data into a learning and a testing sample. The predictive power of the logistic regression model in terms of the c-index (45) using a test sample of 6,395 individual records was 0.71. Twelve preoperative risk factors were found to be significant at a level of $p < 0.01$ by the stepwise logistic regression analysis. Listed in order according to their relative contribution to the total predictive power of the model (given in parenthesis), the risk factors are: prior heart surgery (35.7%), age (13.7%), New York Heart Association functional class (11.0%), peripheral vascular disease (8.8%), cardiomegaly (6.0%), priority of surgery (5.5%), use of intravenous nitroglycerin within 48 hours preceding surgery (4.4), serum creatinine (3.8%), current diuretic use (3.5%), preoperative use of the intra-aortic balloon (3.5%), pulmonary rales (2.1%), and chronic obstructive pulmonary disease (2.0%).

Temporal Trend for all VA Cardiac Surgery

The model with common slope for all hospitals shows that a significant drop of risk-adjusted mortality ($p < 0.01$, LRT) has occurred in the VA system during the last five and one-half years (Figure 1). The model estimates a 4% relative reduction in the adjusted mortality rate per year with a 95% confidence interval of 0.4% to 7.5%. Departures from the overall linear trend using Poisson additive models were found to be significant ($p < 0.01$), showing that the drop in mortality has not been constant over time. The auto-correlation term was found not to be significant when added into the Poisson regression model $\hat{\rho} = -0.069$ and $SE(\hat{\rho}) = 0.05$.

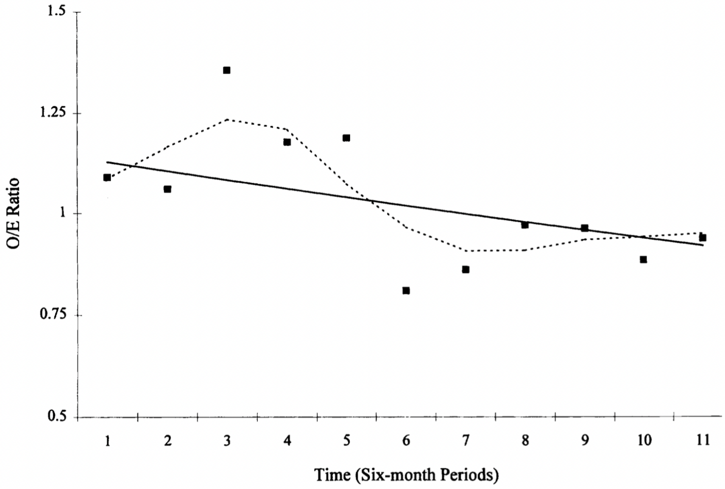


Fig. 1. Risk-adjusted mortality for coronary artery bypass graft patients for 11 6-month periods for an individual hospital (Hospital B) classified as a low outlier based on the overall mean O/E ratio and the downward trend with time. Cross-sectional O/E ratios (dots), trend with time (solid Iine), and mean O/E ratio wlth tlme (dashed line) were estimated by a Poisson regression model

Deviations in Temporal Trends of Individual Hospital O/E Ratios

Results of fitting the model with hospital-specific trends over time, model (1), show that some individual hospitals significantly depart from the overall VA trend ($p < 0.01$, LRT). This result anticipates the presence of high or/and low outlier hospitals based on mean and trend O/E ratio over time. Four hospitals were classified as having a significantly high and one hospital was identified as having a significantly low mean O/E ratio over time respectively (Table 1). Three hospitals were identified as having a significant upward trend and seven hospitals were identified as having a significant downward trend overtime (Table 1).

Only one of the high hospital outliers based on the trend, hospital A, is also a high outlier according to the mean O/E ratio over time (Figure 2). This hospital has a mean O/E ratio of 1.49 with a 95% confidence interval of (1.14, 2.33) and an average increment of risk-adjusted mortality of 42% per year with a 95% confidence interval of (11%, 80%).

One of the low outliers based on trend over time, hospital B, was identified as a low outlier based on mean O/E ratio (Figure 3). This hospital has a mean O/E ratio of 0.47 with a 95% confidence interval of (0.30, 0.74) and a mean adjusted mortality reduction of 41% per year with a 95% confidence interval of (33%, 55%).

Three hospitals were classified as having a constant but high mean O/E ratio over time. For example, hospital C (Figure 4) has a mean O/E ratio of 1.67 (with a 95% confidence interval of 1.28, 2.18), but a mean adjusted mortality reduction of only 1% per year.

To compare the consistency of the TSMO methodology versus the classification of hospitals based on a six-month period, an evaluation of the individual six-month O/E ratios for the last three years was used. This three-year window was chosen to make the comparison more relevant since the TSMO is a longitudinal analysis predicting results based on previous performances.

Hospital A, shown in Figure 2, was found to be a high outlier only two of six times in the past three years based on cross-sectional analysis of O/E ratios. Hospital B, shown in Figure 3, was classified a low outlier four of the last six time periods based on the same analysis. Finally, Hospital C, which was found to have a consistent high mean and stable trend O/E ratio over time (Figure 4), was classified a high hospital outlier only once in the last three years using individual O/E ratio analysis.

To explore the potential correlation of these TSMO findings with the Cardiac Surgeon's Consultants Committee historical activities, the minutes of the Committee meetings were reviewed retrospectively. Hospital A had an intensive history of Committee review level based multiple chart audits and site visits scheduled during this period. Hospital B had no chart audits or site visits initiated by the Committee. Hospital C had received moderate review activity based on multiple chart audits, but no site visits had been scheduled. Although this correlation is inconclusive, there appears to be a positive relationship between the historical level of surveillance and the TSMO findings. More research, however, is needed to conduct a true validation of the TSMO methodology as an independent quality of care assessment tool.

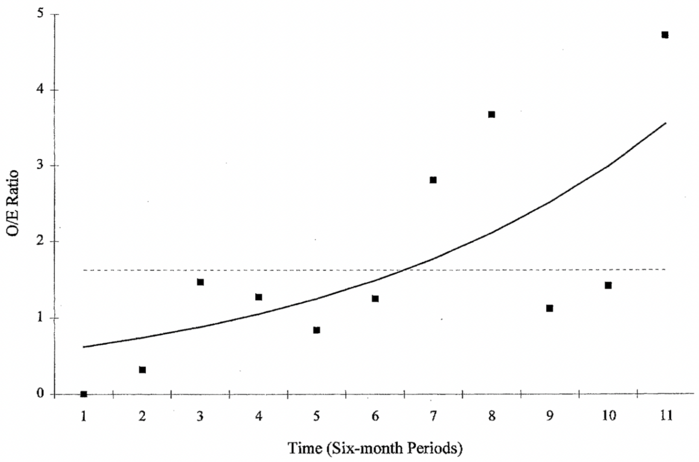


Fig. 2. Rlsk-adjusted mortality for coronary artery bypass graft patients for 11 6-month periods for an individual hospital (Hospital A) classified as a high outlier based on the overall mean observed-to-expected (O/E) ratio and the upward trend with time. Cross-sectional O/E ratios (dots), trend with time (solid line), and mean O/E ratio with time (dashed line) were estimated by a Poisson regression model

5 Discussion

The most important goal of any quality of care model is to improve patient care. An intermediate goal is to be able to validly and reliably define and measure quality of health care in a timely, relevant manner. Although risk-adjusted mortality rates have not been rigorously validated as indicators of hospital or surgeon quality of cardiac surgical care, it seems intuitively correct that risk-adjusted mortality rates should be a more precise measure of quality of hospital care compared with limited adjustment or unadjusted mortality rates (46; 47).

The goal of any quality of care assessment and improvement program is to achieve better patient outcomes. There has been a highly significant ($p < 0.01$) decrease of approximately 4% per year in risk-adjusted mortality rates for CABG procedures performed within the VA during the 5.5 years of the Continuous Improvement in Cardiac Surgery Study. Although we can not exclude the possibility that some or all of this decrease in risk-adjusted mortality rates may be due to technological improvements (e.g., pharmaceutical innovations) in the patient care provided, we believe it is unlikely that the observed reduction is the result of selecting lower risk patients for surgery as we have adjusted mortality for a comprehensive set of risk factors.

The VA hospital system, and the states of New York and Pennsylvania in evaluating provider quality performance have used cross-sectional analyses of risk-adjusted mortality rates confined to brief time periods. However, cross-sectional analysis has several disadvantages including a high variability in classification of outlier status over time for individual providers and a lack of adequate statistical power for low volume providers. For example, a hospital performing on average 150 cases per year and with observed O/E ratios of 2.5 and 2.1 (both viewed as clinically relevant deviations) in two consecutive six-month periods would be classified as a high outlier hospital in the first period, but would fail to reach the significance level established for outlier status in the following period. Although there is no evidence that the O/E ratio of 2.1 is significantly different from 1 if no previous information is used, the pattern based on trend analysis makes a O/E ratio of 2.1 very conclusive evidence that this hospital continues to be a high outlier. Although TSMO was designed for the evaluation of hospital quality of care performance, these issues are likely to be exacerbated for physician/provider quality assessment analyses.

Time Series Monitors of Outcome was developed as an analytical strategy to address the deficiencies identified in cross-sectional analysis of risk-adjusted mortality rates. Inherently, TSMO a longitudinal analysis of risk-adjusted mortality rates over time has face validity. TSMO provides an increase in the consistency with which provider performance is evaluated and an increase in the statistical power to detect quality of care issues at low volume providers.

The ability of TSMO to correlate with independent quality review findings is not known. Although not a perfect validation scheme, preliminary findings of historical committee activities indicate an agreement between TSMO findings and VA clinician judgments to conduct medical chart audits and site visits to investigate potential performance problems. However, TSMO outlier status is not a definitive statement regarding the quality of care, but rather raises a quality question requiring further in-depth review.

The TSMO results indicate that this new classification scheme may assist health policy decision makers and hospital management as a more precise method to detect providers with potential quality of care problems or outstanding quality of care processes and structures that may not be identified using traditional risk-adjusted mortality rates for a single period of time due to the inherent statistical variability in cross-sectional data. TSMO is both computationally simple and policy relevant. Based on results of Table 1, it seems to be more appropriate to identify a hospital as a low outlier if it demonstrates any of the following three combinations of mean

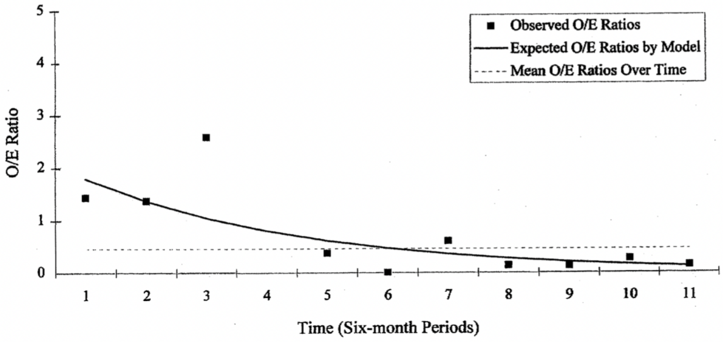


Fig. 3. Risk-adjusted mortality for coronary artery bypass graft patients for 11 6-month periods for an indivldual hospital (Hospittal B) classified as a low outlier based on the overall mean observed-to-expected (O/E) ratio and the downward trend with tlme. Cross-sectional O/E ratios (dots), trend with time (solid line), and mean O/E ratio with tlme (dashed line) were estimated by a Poisson regression model

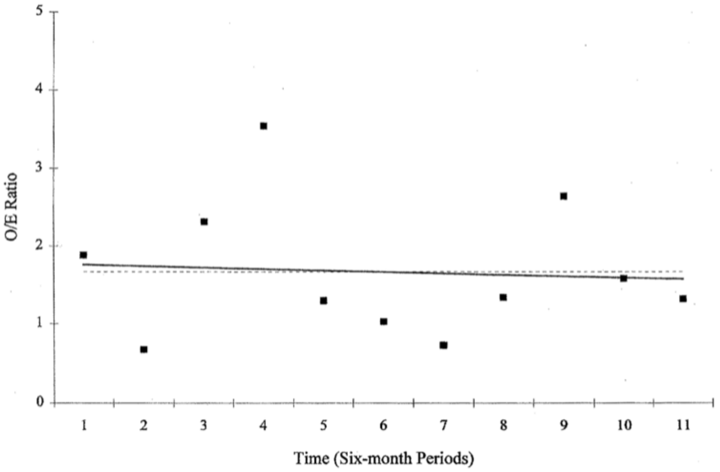


Fig. 4. Risk-adjusted mortality for coronary artery bypass graft patients for 11 6-month periods for an indivldual hospital (Hospittal C) classified as a low outlier based on the overall mean observed-to-expected (O/E) ratio and the downward trend with tlme. Cross-sectional O/E ratios (dots), trend with time (solid line), and mean O/E ratio with tlme (dashed line) were estimated by a Poisson regression model

and trend respectively: 1) low mean with downward trend, 2) low mean with stable trend, or 3) normal mean with downward trend. In all of these cases the hospital is showing a consistent low mortality over time compare with other hospitals. Similarly a high outlier hospital would be defined as one with any of the following combinations of mean and trend respectively: 1) high mean with upward trend, 2) high mean with stable trend, or 3) normal mean with upward trend. In these cases the hospital has a high mortality or is increasing the mortality over time. According to this definition seven hospitals are low outliers (one with a low mean O/E ratio and downward trend and six with normal mean O/E ratios but with a downward trend) and six are high outliers (three with upward O/E ratios, two of which had normal mean O/E ratios and three with stable trends, but persistent high mean O/E ratios).

Low outlier facilities may be viewed as potential "benchmark" facilities from which processes and structures related to high quality of care may be explored. High outlier facilities potentially may have performance problems. Both types of outlier status designations warrant further indepth review of the care provided.

This particular analysis did not show inconsistent classification using mean or trend analysis (i.e., low mean with upward trend or high mean with downward trend). If hospitals were identified as either one of these two cases, it would be important to consider including these facilities in a more intensified quality of care review to understand the reason for these conflicting results. This group of hospitals should be monitored closely for progress or deterioration in subsequent periods.

TSMO is a new tool for identifying relationships between risk-adjusted mortality and time. TSMO appears conceptually to be a quality screening tool applicable to other hospital outcome measures. However, more research is needed to address the relevance, validity, reliability, and clinical implications of the TSMO methodology. TSMO evaluation of O/E ratios provides "outcome stability" to the existing quality screening methods. There is a renewed national emphasis and growing attention placed on quality of patient care issues. Thus, TSMO is a methodology that may improve the reliability and credibility of outcomes to assess the quality of patient care performance.

Bibliography

[1] Goldman, R.L., "The Reliability of Peer Assessments of Quality of Care", Journal of the American Medical Association, Vol 267, No. 7, February 19, 1992, p. 958 - 960.

[2] Rubin, H.R., Rogers, W.H., Kahn, K.L., Rubenstein, L.V., and Brook R.H., "Watching the Doctor-Watchers: How Well Do Peer Review Organization Methods Detect Hospital Quality of Care Problems?", Journal of the American Medical Association, Vol 267, No. 17, May 6, 1992, p. 2349-2354.

[3] Caplan, R.A., Posner, K.L., and Cheney, F.W., "Effect of Outcome on Physician Judgments of Appropriateness of Care", Journal of the American Medical Association, Vol. 265, No. 15, April 17, 1991, p. 1957 - 1960.

[4] Berwick, D.M., "Continuous Improvement as an Ideal in Health Care", New England Journal of Medicine, 1989; 320: 53-56.

[5] McNeil, B.J, Pedersen, S.H., and Gatsonis, C., "Current Issues in Profiling Quality of Care", Inquiry, Fall 1992, 29:298-307.

[6] Nash, D.B., and Goldfield, N. "Information Needs of Purchasers", Chapter 1, Providing Quality Care: The Challenge to Physicians, Goldfield, N. and Nash, D.B. editors, American College of Physicians, 1989.

[7] Neuhauser, D., "Ernest Amory Codman, M.D., and the End Results of Medical Care", International Journal of Technological Assessment in Health Care, 1990, 6:307-325.

[8] Dubois, R.W., Rogers, W.H., Moxley, J.H., Draper, D., Brook, R.H., "Hospital Inpatient Mortality: Is It a Predictor of Quality?", New England Journal of Medicine 1987; 317:1674-1680.

[9] Nevers, R.L. "Defining quality is difficult, but necessary." Healthcare Financial Management Journal, February 1993, p. 18.

[10] Hebel, J.R., Kessler, I.I., Mabuchi, K. and McCarter, R.J., "Assessment of Hospital Performance by Use of Death Rates: A Recent Case History", Journal of the American Medical Association, Vol 248, No. 23, December 17, 1982, p. 3131-3135.

[11] Fleming, S.T., McMahon, L.F., Desharnais, S.I., Chesney, J.D., and Wroblewski, R.T., "The Measurement of Mortality: A Risk-Adjusted Variable Time Window Approach", Medical Care, Vol. 29, No. 9, September 1991, p. 815 -825.

[12] Smith, D.W., Pine, M., Bailey, R.C., Jones, B., Brewster, A., and Krakauer, H., "Using Clinical Variables to Estimate the Risk of Patient Mortality", Medical Care, Vol. 29, No. 11, November, 1991, p. 1108 - 1119.

[13] Park, R.E., Brook, R.H., Kosecoff, J., Keesey, J., Rubenstein, L., Keeler, E., Kahn, K., Rogers, W.H., and Chassin, M.R., "Explaining Variations in Hospital Death Rates: Randomness, Severity of Illness, and Quality of Care", Journal of the American Medical Association, Vol. 264, No. 4, July 25, 1990, p. 484 - 490.

[14] Duckett, S.J., and Kristofferson, S.M., "An Index of Hospital Performance", Medical Care, Vol. XVI, No. 5, May 1978, p. 400 - 407.

[15] Desharnais, S.I., Chesney, J.D., Wroblewski, R.T., Fleming, S.T., and McMahon, L.F., "The Risk-Adjusted Mortality Index: A New Measure of Hospital Performance", Medical Care, Vol. 26, No. 12, December 1988, P. 1129 - 1145.

[16] Kahn, K.L., Brook, R.H., Draper, D., Keeler, E.B., Rubenstein, L.V., Rogers, W.H., and Kosecoff, J., "Interpreting Hospital Mortality Data: How Can We Proceed?", Journal of the American Medical Association, Vol. 260, No. 24, December 23/30, 1988, p. 3625 - 3628.

[17] Blumberg, M.S., "Risk Adjusting Health Care Outcomes: A Methodological Review", Medical Care Review, 43:2 (Fall 1986) p. 352 - 393.

[18] Kelly, J.V. and Hellinger, F.J., "Physician and Hospital Factors Associated with Mortality of Surgical Patients", Medical Care, Vol. 24, No. 9, September, 1986, p. 785 - 959.

[19] Sloan, F.A., Perrin, J.M., and Valvona, J., "In-hospital mortality of surgical patients: Is there an empirical basis for standard setting?" Surgery, Vol. 99, No. 4, April 1986, p. 446 - 453.

[20] Fink, A., Yano, E.M., and Brooks, R.H., "The Condition of the Literature on Differences in Hospital Mortality", Medical Care, Vol. 27, No. 4, April 1989, p. 315 - 336.

[21] Dubois, R.W., Rogers, W.H., Moxley, J.H., Draper, D., and Brook, R.H., "SPECIAL REPORT: Hospital Inpatient Mortality: Is It a Predictor of Quality?", New England Journal of Medicine, Vol. 317, No. 26, December 24, 1987, p. 1674 - 1680.

[22] Dubois, R.W., Brook, R.H., Rogers, W.H., "Adjusted Hospital Death Rates: A Potential Screen for Quality of Medical Care", American Journal of Public Health, September 1987, Vol. 77, No. 9, p. 1162 - 1166.

[23] Hartz, A.J., Krakauer, H., Kuhn, E.M., Young, M., Jacobsen, S.J., Gay, G., Muenz, L., Katzoff, M., Bailey, R.C., and Rimm, A.A., "SPECIAL ARTICLE: Hospital Characteristics and Mortality Rates", New England Journal of Medicine, Vol. 321, No. 25, December 21, 1989, p. 1720 - 1725.

[24] Taulbee, P., "Outcomes Management: Buying Value and Cutting Costs", Business and Health, March 1991.

[25] Hammermeister KE, Johnson R, Marshall G, Grover FL., "Continuous Assessment and Improvement in Quality of Care: A Model from the Department of Veterans Affairs Cardiac Surgery", Annals of Surgery, 1994; 219: 281-290.

[26] Hannan, E.L., Kilburn, H., O'Donnell, J.F., Lukacik, G., Shields, E.P., "Adult Open Heart Surgery in New York State: An Analysis of Risk Factors and Hospital Mortality Rates", Journal of the American Medical Association, Vol 264, No 21, December 5, 1990, p. 27682774.

[27] Office of Health Systems Management, Coronary Artery Bypass Graft Surgery in New York State, New York State Department of Health, December 1992.

[28] The Pennsylvania Health Care Cost Containment Council, Coronary Artery Bypass Graft Surgery: A Technical Report, Harrisburg, PA, November 1992.

[29] Wood, P., "Dealing with Physician-Specific Data Reports: Talking Points for Physicians", Pennsylvania Medicine, January 1993, p.20 - 21.

[30] Iowa Health Data Commission, "Iowa Hospital Resource and Outcome Report: July 1990 through June 1991", April 1992.

[31] Colorado Health Data Commission, "Strategic Plan: 1993-94", September 1993.

[32] Grover, F.L., Hammermeister, K.E., Burchfiel, C., and Cardiac Surgeons of the Department of Veterans Affairs, 'Initial Report of the Veterans Administration Preoperative Risk Assessment Study for Cardiac Surgery', Annals of Thorac Surgery, 1990, 50, 12-28.

[33] Melville, B., "New York Sees Improvement in Cardiac Surgery Outcomes", Report on Medical Guidelines and Outcomes Research, Vol. 3, No. 24, December 23, 1992, p.1-2.

[34] Vilgilante, G.J., Weintraub, W.S., Klein, L.W., Schneider, R.M., Seelaus, P.A.,Parr, G.V.S., Agarwal, J.B., and Helfant, R.H., "Medical and Surgical Survival in Coronary Artery Disease in the 1980s", American Journal of Cardiology, Vol 58, 1986: 926-931.

[35] Project HOPE, "Trends in the Concentration of Six Surgical Procedures under PPS and their Implications for Patient Mortality and Medicare Cost", Technical Report #E-87-08, PROPAC Technical Report Series, June 1988.

[36] Department of Veteran's Affairs Cardiac Surgeons Consultant's Committee, "Criteria and Standards for Monitoring Cardiac Surgery Programs", approved January 1991.

[37] For VA Cardiac Surgeons and Cardiologists: Hammermeister, K.E., Burchfiel, C., Johnson, R., and Grover, F.L. "Identification of Patients at Greatest Risk for Developing Major Complications at Cardiac Surgery," Circulation 1990;82 (Suppl IV):IV380 - IV389.

[38] SAS Institute, Inc., 'SAS/STAT Guide for Personal Computers', Version 6 Edition, Cary, NC, (1985).

[39] Santner TJ and Duffy DE. The Statistical Analysis of Discrete Data. (Pages 34-35). New York, Springer-Verlag, 1989.

[40] McCullagh P and Nelder JA. Generalized Linear Models. Second Edition. London, Chapman and Hall, 1989.

[41] Hastie TJ and Tibshirani RJ. Generalized Additive Models. London, Chapman and Hall, 1990.

[42] Liang KY and Zeger SL. Longitudinal data analysis using generalized linear models. Biometrika, 1986; 73: 13-27.

[43] SAS Institute Inc., SAS Technical Report P-243, SAS/STAT Software: The GENMOD Procedure, Release 6.09, Cary, NC: SAS Institute Inc., 1993.

[44] Marshall G, Grover FL, Henderson WG, Hammermeister KE. Assessment of Predictive Models for Binary Outcomes: An Empirical Approach Using Operative Death from Cardiac Surgery. Statistics in Medicine, 1994, 13: 1501:1511.

[45] Hanley, J.A. and McNeil, B.J., 'The Meaning and Use of the Area under a Receiver Operating Characteristic (ROC) Curve', Radiology, 1982, 143, 29-36.

[46] O'Connor, G.T., Plume, S.K., Olmstead, E.M., et al. "A Regional Prospective Study of In-Hospital Mortality Associated With Coronary Artery Bypass Grafting," Journal of the American Medical Association 1991;266:803-809.

[47] Copeland, G.P., Jones, D., and Walters, M. "POSSUM: A Scoring System for Surgical Audit," British Journal of Surgery 1991;78:356-360.

Article 4.3

A new score for predicting neonatal very low birth-weight mortality risk in the NEOCOSUR South American Network

Guillermo Marshall, Jose L. Tapia, Ivonne D'Apremont, et al.

Pontificia Universidad Católica de Chile

Abstract. To develop and validate a model for very low birth-weight (VLBW) neonatal mortality prediction, based on commonly available data at birth, in 16 neonatal intensive care units (NICU´s) from five South American countries. Prospectively collected bio-demographic data from the NEOCOSUR Network between october 2000 and may 2003 in infants with birth-weight 500 to 1500 gm were employed. A testing sample and cross-validation techniques were used to validate a statistical model for risk of in-hospital mortality. The new risk score was compared with two existing scores by using area under the ROC curve (AUC). The new NEOCOSUR score was highly predictive for in-hospital mortality (AUC = 0.85) and performed better than the CRIB and the NICHD risk models when used in the NEOCOSUR Network. The new score is also well calibrated—it had good predictive capability for in-hospital mortality at all levels of risk (HL test = 11.9, p = 0.85). The new score also performed well when used to predict in hospital neurological and respiratory complications. A new and relatively simple VLBW mortality risk score had a good prediction performance in a South American network population. This is an important tool for comparison purposes among NICU´s. This score may prove to be a better model for application in developing countries.[1]

Keywords: Very low birth weight, Neonates, Infants, Low birth weight.

1 Introduction

Risk adjusted mortality has been used to compare hospital's performance. To accomplish this task, risk models have been developed to adjust raw mortality rates for patient risk differences (1; 2; 3; 4). Development of a mortality risk measure is essential to compare outcomes across NICU´s. Evaluation of medical practices, benchmarking purposes and quality of care comparisons require accurate and reliable risk models. Several neonatal mortality risk scores have been developed, including the Clinical Risk Index for Babies (CRIB) by the International Neonatal Network1, the NICHD score developed by Horbar and colleagues (2), and the Score for Neonatal Acute Physiology (SNAP and SNAPPE-II) by Richardson and colleagues (5; 6). All of these scores have been developed using outcomes from NICU networks in developed countries, where the human and technical resources are very similar with few limitations.

[1] Marshall G, Tapia JL, D'Apremont I, et al. 2005. A new score for predicting neonatal very low birth-weight mortality risk in the NEOCOSUR South American Network. Journal of Perinatology 25 (9), 577-582

The purpose of this study was to develop a neonatal mortality risk score for very- low- birth-weight (VLBW) infants based on variables present at birth, before NICU admission in a multi-center South American Network with diverse mortality rates and resources.

2 Methods

All infants with BW 500g to 1500g born from 10/1/2000 through 5/30/2003 in 16 NEOCOSUR Network participating centers from Argentina, Chile, Paraguay, Peru and Uruguay were included in this study. Biodemographic information and outcome data are prospectively and routinely collected in the NEOCOSUR Network using predefined diagnostic criteria and online data entry. Delivery room deaths were included.

The following risk factors associated with adverse outcome were included in the model: maternal age, birth weight, gestational age, 1 minute Apgar score, major acute life-threatening congenital malformations, sex of the infant, multiple birth, prenatal steroids use, and small for gestation age (defined as lower than the 10th percentile of weight for their gestation age according to a Chilean national growth curve (7)) were studied. Acute life threatening (ALT) congenital malformations included diaphragmatic hernia, major congenital heart disease, intestinal atresia, hydrops, and inborn errors of metabolism, as in the CRIB study. The candidate risk factors were chosen among the ones described in other studies and based on variables present at birth. Univariate association between these infant risk factors and mortality were performed using a simple logistic regression model for categorical variables and a generalized logistic additive model for continuous variables. This model was used to assess the form of the effect of the continuous variables, using non-parametric curve estimation.

A stepwise multiple logistic regression model was used to select the subset of variables that were independently associated with mortality. A significance level of 5% was used to include a variable in the model. With the development of the final model, we were able to estimate the probability of in-hospital mortality for each infant based on prenatal and admission risk factors.

The model was developed using a random sample consisting of 75% of the cases (model sample, n = 1351). The rest of the data was set aside for model validation purposes (test sample, n = 450). A secondary strategy of validation was used using cross-validation techniques (8). We used our final multiple logistic regression model for cross validation, by using it to obtain predicted mortality estimates for each infant (n=1801), based on data from all other infant. This provided an alternative assessment of the predictive capability and calibration of our model.

To assess the predictive capability of the model, the area under the curve (9) (AUC) was calculated using the model sample, the test sample and the total sample using cross-validation techniques. Comparisons with other risk scoring were done using Bootstrap confidence intervals of differences in AUC (10). The calibration of the model was performed using the Hosmer-Lemeshow test (11). The statistical analysis was done using Splus software (12).

The model was also used to predict in-hospital relevant respiratory and neurological complications. Bronchopulmonary dysplasia (BPD) was defined as an oxygen requirement at 28 days of life and chronic radiographic changes (13). Oxygen dependency at 36 weeks postmenstrual age was considered a separate diagnosis The diagnosis of intraventricular hemorrhage (IVH) was made by cranial ultrasonography (this was done at least twice, in the first week of life, and at 3-4 weeks age) or by autopsy and was classified according to Papile et al (14). Periventricular leucomalacia (PVL) was diagnosed by the presence of focal echolucencies on the cranial ultrasound.

Table 1. Comparison of population characteristics between very low birth weight infants survivors and non survivors in relation to factors used to construct the NEOCOSUR risk score. Mean ± SEM or percentage when appropriate.

Factor	Survivor (n=1322)	Non-Survivor (n=479)
Birth weight (grams)	1161 ± 6.7	863 ± 11.9
Gestational Age (weeks)	30.1 ± 0.1	27.0 ± 0.1
1-minute Apgar	6.6 ± 0.1	3.9 ± 0.1
Congenital malformation (1)	0.3%	7.7%
Prenatal steroids use	74%	53%
Mother's Age (years)	28 ± 0.2	26 ± 0.3
Small For Gestation Age (2)	45%	53%
Female gender	52%	46%
Multiple Birth	19%	15%

(1)Acute life threatening

(2) Defined as lower than the 10th percentile of weight for their gestation age (7)

3 Results

The study population included 1801 infants. The mean birth weight was 1081g, and the mean gestational age was 29.2 weeks. Male infants represent the 49.4% of the total population. The mean mortality rate among the different centers was 26.7% with a range of 9.7% to 51.8% among units.

Table 1 shows descriptive statistics comparing survivors and non-survivors. In average, survivors have more birth weight, gestational age, and 1-minute Apgar than non-survivors. Additionally, survivors have less congenital malformations and male infants than non-survivors.

Table 2 displays the univariate associations between in-hospital mortality and prenatal and admission infant characteristics. Birth weight was the best predictor of mortality (AUC = 0.79), followed by gestational age (AUC = 0.77) and 1-minute Apgar score with the same association level (AUC = 0.77). Other factors that were significantly associated with in-hospital mortality were ALT congenital malformations, prenatal steroid use, maternal age and small for gestational age.

The effects of each of the quantitative variables on in-hospital mortality were analyzed using generalized additive logistic regression models. All variables except maternal age showed linear effect, therefore the data were entered into the multiple logistic regression model as they were originally measured, without breaking them into range categories.

When a multiple logistic regression model with a step-wise procedure was used, six factors were statistically significant in order of significance: birth weight, gestational age, 1-minute Apgar score, ALT congenital malformation, antenatal steroid administration, and female gender of the infant. Table 3 shows the coefficients of the multiple logistic regression model, the standard error and the associated odds ratios. These coefficients for the six selected variables in the context of a logistic model constitute the NEOCOSUR score and can be used to calculate predictive risk of VLBW infant mortality. An increase of 100 grams in birth weight reduced the risk of in-hospital mortality by 28%. Similarly, one addition point in the 1-minute Apgar score reduced the risk by 23%. One additional week of gestation age reduced the risk by 12%. ALT congenital malformation

Table 2. Univariate association between in-hospital mortality and pre-admission infant characteristics

Factor	χ^2 test	p-value
Birth weight (grams)	343.6	< 0.01
Gestational Age (weeks)	290.3	< 0.01
1-minute Apgar	289.9	< 0.01
Congenital malformation (1)	63.3	< 0.01
Prenatal steroids use	52.1	< 0.01
Mother's Age (years)	7.8	< 0.01
Small For Gestation Age (2)	4.1	0.04
Female gender	3.2	0.07
Multiple Birth	2.1	0.15

(1)Acute life threatening

(2) Defined as lower than the 10th percentile of weight for their gestation age (7)

Table 3. Factors selected by a stepwise logistic regression model that constitute the NEOCOSUR score for predicting VLBW infant mortality.

Factor	Coefficient	SE	Odds Ratio	95% CI
Constant	8.378	0.99		
Birth weight (grams)	-0.331	0.04	0.72	0.64-0.79
Gestational Age (weeks)	-0.132	0.04	0.88	0.80-0.95
1-minute Apgar	-0.265	0.03	0.77	0.71-0.82
Congenital malformation (1)	3.419	0.64	30.55	8.74-106.77
Prenatal steroids use	-0.302	0.08	0.74	0.58-0.89
Female gender	-0.474	0.16	0.62	0.30-0.93

(1)Acute life threatening

had the largest effect size among all the risk factors–it increases the risk of neonatal mortality by more than 5 times. However, the population attributable risk due to this factor would be relatively small because ALT congenital malformations have an incidence of only 2.2%. Table 4 shows the predictive capability of the NEOCOSUR score as compared to the CRIB and the NICHD scores using the area under the receiver operator curve (AUC). The NEOCOSUR risk score had a high predictive ability (AUC= 0.88) when it was evaluated on the model sample (n = 1351); however, its predictive ability was reduced to 0.84 when it was used to predict mortality in the test sample (n = 450). When the NEOCOSUR score was evaluated in the two combined samples, the AUC was 0.87. As an alternative to the testing sample, the NEOCOSUR score had an AUC of 0.85 when cross-validation techniques were used. The NEOCOSUR had better predictive capacity for in-hospital neonatal mortality than either the CRIB and NICHD scores, or birth weight alone.

The 95% confidence interval for the difference between the AUC of the NEOCOSUR score, using cross-validation result, and the CRIB score was $(0.037 - 0.081)$ and between the AUC of the NEOCOSUR score and the NICHD score was $(0.019 - 0.040)$. With these results we can conclude that the NEOCOSUR score is significantly better predictor of mortality for this population. Figure 1 shows the ROC curve resulting from the NEOCOSUR score using the cross-validated score estimate, the NICHD score, and the CRIB score in all observations (n = 1,801). The NEOCOSUR score had higher sensitivity than the NICHD and CRIB scores, for all

Table 4. Predictive capability of various neonatal mortality risk scores using the area under the ROC curve (AUC), including the NEOCOSUR score.

Risk Score	Sample	AUC
NEOCOSUR	Model	0.88
NEOCOSUR	Test	0.84
NEOCOSUR	Total	0.87
NEOCOSUR	Cross-validation	0.85
CRIB	Total	0.79
NICHD	Total	0.83
Birth Weight	Total	0.79

Table 5. Area under the ROC for adverse outcomes among surviving infants using the NEOCOSUR, CRIB and NICHD scores.

In-hospital Complication	AUC of Risk Score		
	NEOCOSUR	CRIB	NICHD
Severe ICH	0.72	0.77	0.69
PVL (1)	0.69	0.75	0.67
Oxygen 36 weeks	0.77	0.72	0.76
BPD	0.81	0.76	0.80
HIC Grade III-IV	0.72	0.77	0.69

(1) periventricular leucomalacia

levels of specificity. Figure 2 shows the goodness of fit of the NEOCOSUR score by comparing the observed versus estimated mortality rate for each deciles group of risk score. A Hosmer-Lemeshow goodness of fit test (Chi-square = 11.9, degree of freedom = 8, p = 0.85) confirms that the NEOCOSUR score assess mortality rate well at all levels of risk.

When the NEOCOSUR score was used to predict in-hospital neurological and respiratory complications among VLBW infants that survive to hospital discharge (Table 5), there was good predictive ability for BPD at 28 days (AUC = 0.81) and oxygen requirement at 36 weeks postmenstrual age (AUC = 0.77), but slightly lower values for IVH grades III-IV grade (AUC = 0.72) and PVL (AUC = 0.69). When compared with the CRIB and NICHD risk scores, NEOCOSUR score shows similar predictive capability for these morbidities.

4 Discussion

There is a wide variation in VLBW infant mortality rates in South America, as reported previously (15). The reported sample was born in NEOCOSUR Network centers that are university affiliated and belong to the public and private health care systems, but vary in size, population served, and resources, The in-hospital mortality rates ranged from 9.7 to 51.8%. We developed a new relatively simple mortality prediction model for VLBW infants that could be successfully applied soon after birth in an area such as South American that has diverse levels of care and population risks

We tested for commonly available factors that are present before the 1-minute Apgar score, so that prediction would not be influenced by postnatal interventions. We studied several variables that affect mortality, and selected the ones that were most highly associated with mortality.

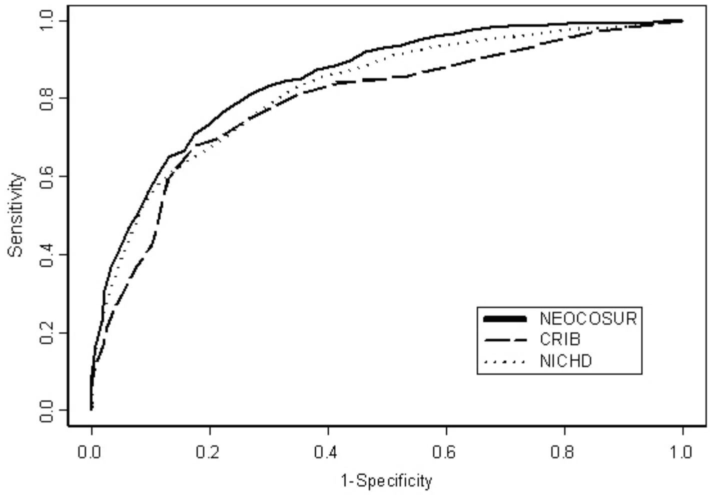


Fig. 1. ROC curves for in-hospital mortality for NEOCOSUR, CRIB, and NICHD risk scores

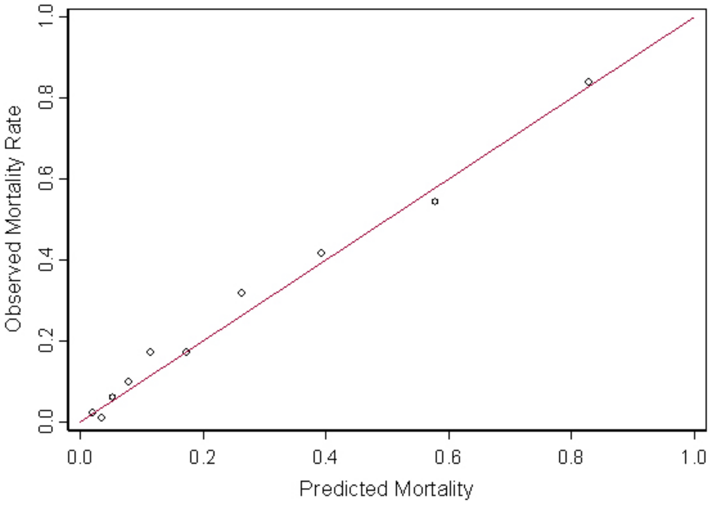


Fig. 2. RObserved mortality rate by decile of predicted mortality for VLBW infants (n=1,801). The predicted mortality was calculated based on a logistic regression model. The dots represent the observed mortality rate of the decile groups and the continuous line represent perfect model calibration.

Birth weight was found to be the variable most predictive of in-hospital mortality, followed by gestational age, and 1-minute Apgar score. Although birth weight is recognized as a major determinant of neonatal mortality, it is inadequate to explain the large variations in neonatal mortality among NICUs16 ; so that a mortality risk score is required.

The model was validated using two alternative methods. First, the model was validated using a test sample not used for model development, and second, the model was validated using cross validation techniques, a leave-one-out resampling method. Both methods consistently showed that the NEOCOSUR score has better predictive capabilities than the CRIB and NICHD scores for in-hospital mortality in this population.

Our model performed better than the CRIB and NICHD models. CRIB was developed in the United Kingdom (n= 1,300; BW ¡ 1500g) and published in 1993. It consists of six items, including birth weight, gestational age, congenital anomalies and three physiological measures obtained during the first 12 hours of age. SNAP was developed in the Unites States (n = 1,643 admissions) and published in 1993, and includes newborns of all birth weight. It consists of 34 items collected in the first 24 hours of admission. We were unable to compare our results with the SNAP because we did not have all the information about the items required. The NICHD score was developed in the United States (n=3603 infants 501-1500g) and was also published in 1993. It consists of five items including birth weight, small for gestational age, black race, male gender and 1-minute Apgar score.

These scores have been replicated with good to excellent performance in most reports (17), although with occasional poor performance and no better than birth weight alone (18; 19). In recent years Pollack et al (20) has published risk models in a cohort of VLBW infants from the Washington DC area. They found that these scores over-predicted mortality indicating a need for frequent recalibration. A need for periodic revalidation of risk models has been addressed in the past (17). There have also been other efforts in individual centers to create their own risk score (21).

Our model differs from the CRIB and SNAP risk scores in that it has fewer variables and the information is collected upon admission, therefore it is less dependant on postnatal interventions. In this regard, our model is similar to that of the NICHD. The main difference between the NICHD score and the NEOCOSUR score is that the NICHD score includes race which does not apply in the South American population and our model uses the Apgar score as a continuous variable. Our score is similar to other risk scores in providing an objective initial mortality risk prognosis, but it does not predict the risk for an individual infant and therefore cannot be used to justify the withdrawal of therapy or to limit care.

This new model is primarily designed to be use for mortality risk prediction; although we also found that it may predict serious in-hospital respiratory and neurological complications. However, this last finding is expected since BPD and IVH are closely related to infant's birth weight and gestational age. Its usefulness in predicting long term outcome has not been tested. Tests of the CRIB score suggest that it is not a reliable tool for predicting neurodevelopmental outcome (22).

An obvious limitation of the NECOSUR risk score is that only applies to VLBW infants, under 1500g of birth weight. We conclude that this new and relatively simple neonatal VLBW infant mortality risk score has good predictive performance in a multicenter South American population, and is an important tool for comparison purposes among NICUs. Based on its simplicity and good performance in a diverse population setting, we speculate that this risk score may prove to be a better model for application in developing countries.

Bibliography

[1] The International Network Group. The CRIB (clinical risk index for babies) score: a tool for assessing initial neonatal risk and comparing performance of neonatal intensive care units. Lancet,1993; 342:193-198.

[2] Horbar JD, Onstad L, Wright E, The National institute of Child Healh and Human Development Neonatal Research Network. Predicting mortality risk for infants weighing 501 to 1500 grams at birth: The National Institutes of Health Neonatal Research Network report. Crit Care Med 1993; 21:12-17.

[3] Hammermeister KE, Johnson R, Marshall G, and Grover FL. Continuous Assessment and Improvement in Quality of Care. Annals of Surgery. 1994; 219:281-290.

[4] Marshall G, Shroyer ALW, Grover FL, and Hammermeister KE. Time series monitor of outcomes: A new dimension of measuring quality of care. Medical Care, 1998; 36: 348-356.

[5] Richardson D K, Gray Je, McCormick MC et al. Score for Neonatal Acute Physiology (SNAP): Validation of a new physiology-based severity of illness. Pediatrics 1993; 91:617-623.

[6] Richardson DK, Corcoran JD, Escobar GJ, Lee SK, for the Canadian NICU Network. SNAP-II and SNAAPPE-II: Simplified newborn illness severity and mortality risk scores. J Pediatr 2001; 138:92-100.

[7] Juez G, Lucero E, Ventura-Juncada P, Gonzalez H, Tapia JL, Winter A. Crecimiento intrauterino en recien nacidos de clase media. Rev Chil Pediatr 1989; 60:198-202.

[8] Efron B. Estimating the error rate of a prediction rule: Improvement on cross-validation. J. of the American Statistical Association; 1983; 78:316-331.

[9] Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 1982; 143:29-36.

[10] DiCiccio, TJ and Efron B. More accurate confidence intervals in exponential families. Biometrika, 1992; 79:231-245.

[11] Hosmer DW and Lemeshow S. Applied Logistic Regression. Second Edition. New York, NY: John Wiley and Sons; 2000.

[12] Chambers JM and Hastie TJ. Statistical Models in S. New York: Chapman and Hall; 1992.

[13] Bancalari A, Abdenour G, Feller R, Gannon J. Bronchopulmonary dysplasia: Clinical presentation. J Pediatr 1979; 95: 819-822.

[14] Papile L, Bursten J. Incidence and evolution of subependimal and intraventricular hemorrhage. A study of infants with birth weights less than 1500 g. J Pediatr 1978; 92:529-534.

[15] Grupo Colaborativo Neocosur. Very Low Birth Weight Infants outcome in 11 Southamerican NICU´s. J Perinat 2002, 22:2-7.

[16] Richardson DK, Phibbs CS, Gray JE et al. Birth weight and severity illness: independent predictors of neonatal mortality. Pediatrics 1993; 91:969-975.

[17] Richardson DK, Tarnow-Mordi WO, Escobar GJ. Neonatal risk scoring systems. Clin Perinat 1998;25:591-611.

[18] de Courcy-Wheeler RH, Wolfe RH, Fitzgerald A, et al. Use of the CRIB core in prediction of neonatal morbidity and mortality. Arch Dis Child Fetal Neonatal Ed 1996; 74:F79-80.

[19] Khanna R, Taneja V, Singh SK, et al. The clinical risk index for babies (CRIB) score in India. Indian J Pediatr 2002; 69:957-960.

[20] Pollack MM, Koch MA, Bartel DA et al. A Comparison of Neonatal Mortality Risk Prediction Models in Very Low Birth Weight Infants. Pediatrics 2000; 105:1051-1057.

[21] Maier RF, Rey M, Metze BC et al. Comparison of mortality risk: a score for very low birthweight infants. Arch Dis Child 1997; F146-F151.

[22] Lago P, Freato F, Bettiol T et al. Is the CRIB score a valid tool in predicting neurodevelopmental outcome in extremely low birth weight infants? Biol Neonate 1999; 76:220-227.

Article 4.4

Assessment of Predictive Models for Binary Outcome

Guillermo Marshall, Frederick L. Grover, William G. Henderson and Karl E. Hammermeister

University of Colorado Health Sciences Center,
Department of Veterans Affairs Medical Center, and
Department of Veterans Affairs Cooperative Studies Program

Abstract. Predictive models in medical research have gained popularity among physicians as an important tool in medical decision making. Eight methodological strategies for creating predictive models are compared in a large, complex data base consisting of preoperative risk and operative outcome data on 12,712 patients undergoing coronary artery bypass grafting and entered into the Department of Veterans Affairs Cardiac Surgery Risk Assessment Program between April 1987 and March 1990. The models under consideration were developed to predict operative death (any death within 30 days following the surgical procedure or later if the result of a perioperative complication). The two strategies with the best predictive power among the eight examined were stepwise logistic regression alone and data reduction by cluster analysis combined with clinical judgment followed by a logistic regression model. The additive model based on unadjusted relative risks, the model based on Bayes' Theorem, and the logistic model using all candidate variables were good alternatives. Whether or not we imputed values did not have a significant impact on the predictive power of the models.[1]

Keywords: Logistic Regression, Cluster Analysis, Principal Components, Bayes' Theorem, Classification Trees, ROC Curve

1 Introduction

The development of techniques and strategies to determine the best model to predict an outcome in a large data set with a large number of potential predictive variables has become an increasingly important topic in the last few years. While a great deal of effort has been made in improving the existing methodologies to handle this problem (1; 2; 3; 4; 5; 6), there have been relatively few studies (7; 8; 9) to determine which of these available techniques provides the greatest predictive power for dichotomous outcomes.

In this paper we assess the predictive power of eight methodological strategies to develop multivariable models that discriminate between death (any death within 30 days following the surgical procedure or later if the result of a perioperative complication) and survival in patients undergoing cardiac surgery.

Harrell et al. (7) proposed a general c-index of predictive discrimination to measure the ability of a model to predict survival of patients having coronary artery disease. The index is a measure of concordance between the predicted and the observed outcome, and can be applied to different statistical models. In particular, for models with binary outcomes such as logistic

[1] Marshall G, Grover FL, Henderson WG and Hammermeister KE. 1994. Assessment of Predictive Models for Binary Outcome. Statistics in Medicine 13 (15), 1501-1511

regression and others included in this study, the c index reduces to the area under a receiver operator characteristic (ROC) curve as described by Hanley and McNeil (10)

Harrell et al. (7) stated that both reliability and discrimination should be considered in assessing predictive accuracy. Reliability is commonly tested by dividing the test population into subgroups according to predicted outcome and comparing observed against predicted outcome for each subgroup. Because reliability is difficult to quantify since there are so many ways to subdivide patients, we will concentrate in this paper on predictive discrimination. Since both clinical decisions and quality assessment at the patient level depend on the ability to predict outcomes for the individual patient, it is important to assess how well the model discriminates between favorable and poor outcomes. Variations of receiver operator characteristic (ROC) curves are helpful in measuring discriminatory ability (10) . For both physicians and patients, it is also important to have an instrument to assess the absolute risk of operative mortality.

Harrell et al. (7) evaluated stepwise selection, principal component analysis and clinical scores based on cluster analysis as data reduction methods to reduce the number of variables entered into a multiple Cox's regression model. Harrell and colleagues (8) also evaluated two additional approaches; classification and regression trees and a sickness score based on subjective medical judgment. Cook and Goldman9 discussed the advantages and disadvantages of classification trees (recursive partitioning). Edwards et al. (1; 2) applied Bayes' Theorem to obtain posterior probabilities of a disease given a set of patient characteristics or symptoms. Parsonnet et al. (3) proposed an additive model based on unadjusted odds ratios.

In this work we assess the predictive accuracy for the prediction of operative death following cardiac surgery for the following eight models or strategies:

1. Stepwise logistic regression analysis,
2. Logistic regression using all candidate variables,
3. Data reduction by cluster analysis of the predictor variables and clinical judgment followed by stepwise logistic regression,
4. Data reduction by principal components analysis followed by a logistic regression model,
5. A subjectively created 'sickness score',
6. A model based on Bayes' theorem,
7. An additive model based on unadjusted relative risk, and
8. A classification tree model.

The authors have intentionally omitted some classical statistical techniques such as linear discriminant analysis as well as some nonparametric alternatives. Some of these methods have been excluded with the belief that they will not perform better than logistic regression due to the high degree of nonnormality of the data. Some new modelling strategies have also been excluded such as generalized additive models, multivariate adaptive regression splines and neural networks, but they will be the subject of future reports.

2 The Data

During the first three years (April 1987 through March 1990) of operation of the Department of Veterans Affairs Cardiac Surgery Risk Assessment Program11, we have received 15,444 data forms on patients undergoing cardiac surgery from 44 VA medical centers. Of these, 12,712 underwent coronary artery bypass grafting as the primary procedure, 2,326 underwent valve surgery, and 406 underwent other procedures. The initial efforts at assessing predictive accuracy were confined to the 12,712 patients undergoing coronary artery bypass grafting, because of the larger population size and greater homogeneity of the patient population.

We randomly divided this population into a learning sample (used to create the predictive model) and a test sample (used to test the predictive model) of approximately equal sizes using a pseudo-random number generating function. The sample size and the number of deaths in the two samples were not forced to be equal to prevent arbitrary similarity in outcome distribution. Nevertheless, the sample size, the number of events, and the percent operative mortality were very similar in the two samples as shown in Table 1.

Table 1. Sample Size, Number of Events and Percent Operative Mortality in the Learning and Test Samples and in the Total Population.

Sample	Size	Events	Percent
Learning	6,317	285	4.51%
Test	6,395	297	4.64%
Total	12,712	582	4.58%

From this large data set we selected 43 preoperative variables thought to be associated with operative mortality. Of these, 10 variables with more than 20% of data missing were excluded. The remaining 33 variables and their percent of data missing are shown in Table 2.

The number of patients in the learning data set provides the power to detect an odds ratio of 1.2 in a logistic regression model with various predictor variables. This sample size was calculated using the method described by Hsieh12, assuming a multiple correlation R=0.5 among the predictor variables, statistical power of 90% and a significance level of 5%.

3 Stepwise Variable Selection

Stepwise procedures have been used for many years as a strategy to select variables in statistical models such as multiple linear regression, linear discriminant analysis, multiple logistic regression and Cox's regression models. The statistical algorithm searches for the best model that combines statistical accuracy with parsimony. The algorithm should have criteria to prevent over fitting the data through the inclusion of irrelevant variables in the model. The ideal situation would be to fit all possible models and select the best of them based on a classical measure, such as Akaike's information criterion (AIC). However, as the number of predictor variables increases, the number of possible models increases dramatically. For example, with 10 predictor variables there are 1,024 possible models, but this number increases to 4.3 billion for 33 explanatory variables. Fitting all possible models becomes impractical with more than six potential predictor variables.

Stepwise variable selection is a practical alternative to examining all possible models. Even though an optimizing algorithm is used, finding the best model cannot be guaranteed. The criteria to include or delete variables at each step should be similar to the criteria used to select the best model from all possible models. It should be noted that multiple comparisons is a problem inherent to any selection process searching for the best model. Lawless and Singhal13 proposed an innovative method of screening for nonnormal models that has been implemented in some logistic regression programs14.

Table 2. Variable name with less than 20% missing values in the learning sample used to assess the predictive models, description of the variables and the percentage of missing values in the learning sample.

Variable	Description	Missing
AGE	Age (in years)	0.2 %
ANGIOP	Angioplasty =¡ 7days of Surg *	1.5 %
BSA	Body Surface Area (square meters)	2.5 %
CHF	Congestive Heart Failure *	2.3 %
CM	Cardiomegaly (x-ray) *	4.1%
COPD	Chronic Obstructive Pulmonary Disease *	3.3 %
CR	Creatinine (mg/dl)	14.2 %
CURCAB	Current Calcium Channel Blocker Use *	5.9 %
CURDIG	Current Digoxin Use *	5.6 %
CURDIUR	Current Diuretic Use *	5.5 %
CURRBB	Current Beta Blocker Use *	5.9 %
CURRSMOK	Current Smoker *	5.3 %
CVD	Cerebral Vascular Disease *	5.8 %
CXS	Circumflex (% Stenosis)	13.9 %
DIABETES	Diabetes *	5.6 %
ECGLVH	ECG LVH *	7.7 %
EVERSMOK	Ever Smoker *	5.3 %
EXANG	Exertional Angina *	5.6 %
HTN	Hypertension *	5.6 %
IVNTG	Intravenous Nitroglycerine Preoperatively *	5.6 %
LAD	Left Anterior Descending (% Stenosis)	9.4 %
NYHAFC	NY Heart Association Functional Classification	6.2 %
OLDMI	Old Myocardial Infarction (¿30 days) *	4.9 %
PRIORHS	Prior Heart Surgery *	0.3 %
PRIORITY	Priority of Surgery (elective,urgent,emergent)	4.2 %
PROPIABP	Preoperative Intra-aortic Balloon Pump *	1.9 %
PULMR	Pulmonary Rales *	5.9 %
PVD	Peripheral Vascular Disease *	5.5 %
RECMI	Recent MI (=¡30 days) *	3.0 %
RESTANG	Unstable Angina *	6.4 %
RESTSTD	Resting ST-segment Depression *	7.3 %
RCHS	Right Coronary Artery (% Stenosis)	11.4%
SEX	Sex (0=Male,1=Female)	0.2%

* (1=Present, 0=Absent)

Beginning with the 33 variables (Table 2) we used the backward selection procedure available in PROC LOGISTIC14 to exclude variables that do not meet the significance level $p < 0.05$. The final model included the following six variables: prior heart surgery, priority of surgery, New York Heart Association functional class, pulmonary rales, age, and peripheral vascular disease, listed in order of importance according to their contribution to the total chi-square of the model. Subjects with missing values in any of the 33 variables were excluded from the selection process. For the purpose of estimating the final model, only observations having missing values in the six variables listed above (13%) were excluded from the logistic regression.

As an alternative model, we fit a logistic model with all 33 candidate variables. Subjects having missing values in one of the 33 variables, representing a total of 38% of the observations, were excluded from the analysis.

4 Cluster Analysis and Clinical Judgment

Cluster analysis is a statistical technique that searches for natural clusters of observations in a multi-dimensional space generated by a data set. The criteria used to form the clusters are based on the multi-dimensional distance. Observations having the shortest distance to the center of the cluster are assigned to that cluster.

A similar statistical concept has been used to create clusters of variables, but using multiple correlation instead of distance as criteria to form the clusters. Variables are assigned to the cluster having the highest multiple correlation with the remaining variables in the same cluster. According to this procedure, the clusters are formed to maximize the within-cluster correlation and minimize the between-cluster correlation.

The output from the analysis is a score for each cluster, which is the linear combination of the variables that best explain the variance of the variables within a cluster. The resulting coefficients associated with each variable in the cluster, and the variables selected to form the cluster, are not obtained at this point using their association with the response variable. Our goal was to reduce the number of variables by approximately 75% using cluster analysis. From the 33 variables listed in Table 2, we constrained the computer program to produce no more than eight clusters. The algorithm used in this study was the combination of principal components with orthoblique rotation and the nearest component sorting for the iterative reassignment of the variables to the clusters. This method is implemented in the SAS procedure VARCLUS14.

Although most of the clusters obtained had some clinical relevance, we used clinical judgment for a limited rearrangement of the clusters and the weights of the variables to maximize clinical meaning. This process consisted of eliminating all single variables clinically not related with the remaining variables in the cluster or due to low correlation with the cluster. Among the variables excluded from the clusters were age and prior heart surgery. The coefficients of the scores associated with each cluster were slightly modified to increase clinical interpretation. The resulting scores were entered into a logistic regression analysis with some individual variables excluded from the cluster analysis, as we mentioned before. Using the backward selection method, four of the eight scores were found to be significantly associated with operative mortality in addition to age and prior heart surgery acting individually. The selected scores were (Table 2):

$$\begin{aligned}
CHF_{Score} =& 2.0 \cdot CHF + 1.5 \cdot CURDIG + 1.5 \cdot CURDIUR + 1.0 \cdot PULMR,\\
IschemiaI_{Score} =& 2.0 \cdot IVNTG + 1.5 \cdot RESTSTD + 1.5 \cdot PROPIABP +\\
& 1.5 \cdot RECMI + 2.0 \cdot PRIORITY,\\
PVD_{Score} =& PVD + CVD,\\
IschemiaII_{Score} =& 2.0 \cdot RESTANG + 2.0 \cdot NYHAFC + 1.0 \cdot EXANG.
\end{aligned}$$

In the process of selecting the scores and the final model, subjects having missing values in one of the variables involved in the analysis (19%) were excluded from the analysis.

5 Principal Components Analysis

Principal components analysis is a multivariate data reduction technique similar to factor analysis. Principal components analysis selects the linear combination of the variables that best captures the variability of the data in a multi-dimensional space. The resulting scores are linear combinations of the 33 original variables which are selected to be uncorrelated with each other. Three major limitations of principal components analysis are that they are difficult to interpret, they require the collection of data for all 33 of the original variables, and that the principal components score is a missing value if one or more of the 33 original variables is missing. Similar to cluster analysis, we selected the first eight principal components or scores. The resulting scores were included in a stepwise logistic regression that selected a model with five scores. The principal components selected for the model were the first, the second, the fourth, the sixth and the eighth principal components. Subjects having missing values in one of the 33 variables, representing a total of 38% of the observations, were excluded from this analysis.

6 A Sickness Score

A completely subjective sickness score was proposed by a panel of two experienced cardiac surgeons and a cardiologist. One point was added to the score for each of the following conditions or characteristics present: age greater than 70 years, New York Heart Association functional classification equal to IV, unstable angina, peripheral vascular disease, emergent priority of surgery, chronic obstructive pulmonary disease present, and LVEF between 0.25 and 0.34. Two points were added to the score if the patient had prior heart surgery and/or if LVEF was less than 0.25. The resulting score could take a minimum value of 0 and a maximum value of 10. Subjects having missing values in one of the variables used to create the sickness score (24%) were excluded from the computation of the c-index.

7 Bayes' Theorem and Posterior Probability

Edwards et al.1,2 proposed to use the theorem of Bayes as a tool to predict outcomes for the individual patient. The model computes the posterior probability of an operative death given a set of patient characteristics or symptoms.

Assume a set of m patient's characteristics (signs, symptoms or laboratory values) $\mathbf{s} = \{s_1, s_2, ..., s_m\}$, where $s_1, s_2, ..., s_m$ represent indicators of the presence of those characteristics ($s_i = 1$, otherwise $s_i = 0$). Then, the posterior probability of an operative death, $P\{Death|S\}$, can be calculated as

$$P\{Death|\mathbf{s}\} = \frac{P\{\mathbf{s}|Death\}P\{Death\}}{P\{\mathbf{s}|Death\}P\{Death\} + P\{\mathbf{s}|Survival\}P\{Survival\}} \tag{1}$$

where $P\{Death\}$ is the prior probability of operative death, and $P\{Survival\} = 1 - P\{Death\}$. The values for $P\{s|Death\}$ and $P\{s|Survival\}$ can be obtained from the data assuming conditional independence, that is,

$$P\{\mathbf{s}|Death\} = P\{s_1|Death\} \times P\{s_2|Death\} \times \cdots \times P\{s_m|Death\}. \tag{2}$$

$P\{s_i = 1|Death\}$ can be estimated from the observed proportion who have the characteristic present in the subsample of all operative deaths. The variables or characteristics, S=s1, s2, ...,sm, were selected from the 33 predictor variables shown in Table 2 when a significant association with the outcome was present using Pearson's chi-square test or a two independent samples t-test ($p < 0.01$). A total of 24 characteristics were selected: congestive heart failure, current digoxin use, current diuretic use, intravenous nitroglycerine preoperatively, resting ST-segment depression, preoperative intra-aortic balloon pump, recent myocardial infarction, peripheral vascular disease, cerebral vascular disease, unstable angina, ECG LVH, cardiomegaly, old myocardial infarction, pulmonary rales, prior heart surgery, chronic obstructive pulmonary disease , urgent priority of surgery, emergent priority of surgery, New York Heart Association functional classification (NYHAFC) II, NYHAFC III, NYHAFC IV, age between 51 to 60 years old, age between 61 to 70 years old and age older than 70 years. Subjects having missing values in one of the symptoms si (20%) were excluded from the analysis.

8 Additive Model

Parsonnet et al.3 proposed an additive model to stratify patients into five groups of increasing risk of death after cardiac surgery. The model is formed by adding the odds ratios of the selected risk factors obtained from univariate models. The most important criteria to select the risk factors are: 1) The factors must have a significant association with the outcome at the univariate level, 2) The factors must be as free as possible from subjectivity and bias, and 3) The factors or variables must be simple and direct, avoiding the use of compound scores.

The additive model used in the present study differs from Parsonnet's model in using relative risks instead of odds ratios. Although in general the relative risk is a more precise measure of risk when it is possible to calculate, odds ratios and relative risks are very similar in our case since the probability of operative death is rather small.

The additive model can be represented as

$$Risk(\mathbf{s}) = r_1 \times s_1 + r_2 \times s_2 + ... + r_m \times s_m \tag{3}$$

where $r_i = RR_i - 1$, RRi_i is the relative risk for the ith factor and si as before is the indicator of presence of the ith risk factor. In cases of quantitative variables such as age, appropriate categories can be found to categorize the values if they satisfy the inclusion criteria mentioned above.

The same criteria used in the Bayes model were used to select the risk factors for the additive model. A total of 24 characteristics from 19 risk factors were included in the additive model as they were listed above. Subjects having missing values in one of the symptoms si (20%) were excluded from the analysis.

9 Classification Trees

Classification trees4, or recursive partitioning, is a nonparametric classification technique that systematically reduces the sample into subgroups with common risk characteristics according to a decision tree structure. In the process of building a tree, a group of patients is partitioned in two subgroups according to a split in one risk factor or the linear combination of more than one risk factor. In each step, the split with best predictive power is chosen among all possible splits in the set of risk factors. The size of the tree is selected by reducing the cost of misclassification in a portion of the sample reserved for that task or using crossvalidation techniques.

One of the major advantages of recursive partitioning compared with more classical approaches is that it allows nonlinear relationships between predictive factors and outcomes. However, when the majority of the risk factors are dichotomous variables representing the presence or absence of the risk factor, as in our data set (Table 2), this advantage tends to be minimal. A second important advantage of classification trees is its method of handling missing values by surrogate splits. Every titular split has associated with it one or more surrogate splits that can be used when the variable used in the titular split is missing. The surrogate splits are selected among all possible splits for having the maximum association with the titular split in an 2 x 2 contingency table.

Classification trees were fit using the CART computer program and Gini's construction rule4. Due to the large size of the learning sample, a test sample equivalent to 1/3 of the total learning sample was used instead of crossvalidation to determine the size of the tree. A total of 11 terminal subgroups were created in the final tree.

10 Comparative Results

The predictive power of each of these strategies was evaluated in terms of the area under the ROC curve or c index. A value of c equal to 0.5 is equivalent to the prediction of an outcome based on a random function with uniform probability of event. A value of c greater than 0.5 means a discrimination function better than purely random selection, and a value of 1.0 represents an instrument with perfect prediction capability.

Table 3 shows the different models used, the number of factors, variables and splits selected by the model, the c index for the learning sample, and the c index and its standard error for the test sample. As expected, the true predictive power of the models is generally over-estimated in the learning sample compared to the test sample which provides a more unbiased estimate. The over-estimation in the predictive power in the learning sample is largest in the model using classification trees. In contrast, there was no over-estimation in the learning sample with the sickness score model, since it was not created using the data in the learning sample. The similarity of the c index in the learning and test sample for this model can be viewed as a measure of validation of the learning sample.

Logistic regression with stepwise selection and cluster analysis followed by logistic regression are the two best analytic strategies for predicting operative death according to the c index.

Table 3. Results of assessment of the different predictive models, the number of parameters involved in the model, the c index of the learning sample, the percentage of observations in the learning sample excluded due to missing values, the c index in the test sample, and the standard error of the c index

Model	Number of	Learning Sample		Testing Sample	
	Parameters	c-index	% missing	c-index	St. Error
Cluster Analysis	6	0.733	19%	0.711	0.019
Stepwise Logistic Regression	6	0.739	13%	0.710	0.019
Additive Model	24	0.718	20%	0.697	0.021
Bayesian Analysis	24	0.713	20%	0.695	0.021
Logistic Regression	33	0.749	38%	0.694	0.022
Principal Components	5	0.700	38%	0.690	0.022
Sickness Score	8	0.674	24%	0.678	0.020
Classification Trees	11	0.716	0%	0.655	0.017

An advantage for the stepwise selection is the smaller proportion of subjects lost due to missing values compared with cluster analysis. On the other hand, cluster analysis with clinical judgment can be more interpretable than ordinary regression. A small change in the data can change the list of individual variables selected in a stepwise regression analysis, but this is less likely to affect the construction and selection of the cluster indexes.

It is also important to consider other criteria in comparing these models. One of the most important secondary parameters is the pathophysiologic insight that can be obtained from the process of modelling an outcome as a function of many factors. In this sense, stepwise logistic regression and cluster analysis allow greater ease in clinical interpretation than the other models.

Table 4. Results of assessment of the different predictive models in a subsample without missing values, the number of parameters involved in the model, the c index of the learning sample, the c index in the test sample, and the standard error of the c index.

Model	Number of	Learning Sample	Testing Sample	
	Parameters	c-index	c-index	St. Error
Additive Model	24	0.720	0.712	0.021
Logistic Regression	33	0.744	0.711	0.022
Stepwise Logistic Regression	6	0.730	0.710	0.021
Cluster Analysis	6	0.703	0.705	0.021
Bayesian Analysis	24	0.706	0.705	0.021
Classification Trees	6	0.681	0.663	0.021
Sickness Score	8	0.636	0.651	0.021
Principal Components	4	0.704	0.636	0.021

In the process of comparing these models, the effect of missing values can have an important impact on the results shown in Table 3. An ideal situation would be to compare the models under different patterns of missing values. As an alternative to that, Table 3 was reproduced using a sub sample with no missing data in all of the 33 variables shown in Table 2. The results of these analyses are shown in Table 4; now Bayesian analysis, the additive model and the logistic model using all candidate variables have similar predictive power to stepwise logistic regression and cluster analysis.

Table 5. Results of assessment of the different predictive models using MISSGEN to impute missing values, the number of parameters involved in the model, the c index of the learning sample, the c index in the test sample, and the standard error of the c index.

Model	Number of	Learning Sample	Testing Sample	
	Parameters	c-index	c-index	St. Error
Stepwise Logistic Regression	7	0.745	0.710	0.021
Logistic Regression	33	0.751	0.705	0.021
Additive Model	19	0.709	0.700	0.020
Bayesian Analysis	19	0.700	0.694	0.020
Principal Components	4	0.698	0.688	0.020
Cluster Analysis	6	0.714	0.682	0.021
Sickness Score	8	0.672	0.678	0.020

The effect of missing values is also evaluated for seven modelling strategies (excluding classification trees) using imputation of missing values (Table 5) by multiple regressions with the SAS (Statistical Analysis System) macro MISSGEN15. The results do not differ from Table 3 except in the significant reduction of the c index in cluster analysis in the test sample.

11 Discussion

In this paper we compared the predictive capability of eight different methodologies to develop models to predict operative mortality in 12,712 patients undergoing coronary artery bypass surgery at 44 Department of Veterans Affairs Medical Centers over a three-year period. The strengths of this study are that the data were prospectively collected on a large number of patients using a standard data form with uniform definitions for most of the variables.

A potential weakness is that the data were incomplete, because the study did not fund data managers at each medical center to collect the data. On other hand, missing data is a fact of life in nearly all studies of this size, furthermore it gave us the opportunity to study the effect of imputation of missing values on predictive power.

The results of this study indicate that stepwise logistic regression alone or data reduction by cluster analysis of predictor variables with some rearrangement of variables using clinical judgment followed by stepwise logistic regression had the best predictive power among the eight techniques tried. The additive model, the model based on Bayes' Theorem, the logistic model using all candidate variables, and the principal components analysis followed by stepwise logistic regression had c indices that were somewhat smaller than those of the first two methods. The c indices for the sickness score and the classification trees indicated the least predictive power.

Although the c index was relatively high for stepwise logistic regression, the ratio of number of deaths to number of predictor variables in our study was relatively high (297:33 = 9:1). Harrell, et al.8 compared the performance of stepwise logistic regression, a sickness score, principal components analysis, and recursive partitioning using 25 variables to predict complete remission from treatment of 334 patients with non-Hodgkin's lymphoma. In a small training sample of 110 patients with 50 complete remissions, they found principal components and the sickness score to have the highest c indices (0.67 and 0.66, respectively), recursive partitioning to have an intermediate c index (0.61), and stepwise logistic regression to have the smallest c index (0.58). However, when they used a larger training sample of 224 patients with 102 complete remissions, stepwise logistic regression (0.67), the sickness score (0.64), and principal components (0.68) had similar c indices, and recursive partitioning was definitely inferior (0.56). It is interesting that stepwise logistic regression performed relatively well, even though the event-to-variable ratio was only 4:1. The latter results agree with the results of our study. In contrast, Cook and Goldman9 found recursive partitioning to have a larger area under the ROC curve (c index) than logistic regression in using 50 variables for predicting myocardial infarction in 482 patients presenting with chest pain to the emergency room at Yale - New Haven Hospital. However, they did not state the event-to-variable ratio in developing their predictive models which, if it were much less than 10 to 1, could have adversely affected the performance of logistic regression.

When reduction of the number of independent variables prior to regression modelling is important, cluster analysis on the variables or principal components analysis may be used. Our study showed that there was little difference in the predictive power of the final model when either preliminary data reduction method was used. The advantage of principal components analysis is that it is a completely objective method which is invariant to application by different researchers. The disadvantages are that the values of all variables must be present in order to calculate the principal components, and the principal components are difficult to interpret. The same problem is applicable to the logistic regression using all candidate variables. Cluster analysis of the variables, with clinical adjustment of the cluster and weights, followed by logistic regression analysis resulted in the highest predictive power. The advantage of this method is that the clusters of variables have clinical meaning. In our study, of the clusters selected in the regression analysis, two were assessments of ischemia, one was an assessment of congestive heart failure, one was an assessment of peripheral vascular disease, and two clusters were single variables (age and prior heart surgery). The disadvantage of this method is that it is dependent in some degree on clinical judgment and thus may not be replicated exactly by independent investigators even when using the same data set.

Although Parsonnet's additive model seems to perform well in terms of predictive power, it is clearly inappropriate for absolute risk assessment. The use of log odds ratios is recommended. This should not reduce the predictive power of this additive model but can allow the assessment of absolute risk of surgical death. The Bayes' theorem model also shows a relatively high predictive power compared with the other seven models. Although the assumption of conditional independence may be seen here as an important limitation, conditional independence models are widely used in the analysis of categorical data.

The imputation of missing values using the SAS macro MISSGEN does not affect the variable selection process in the two techniques considered. The predictive power remained the same in stepwise logistic regression, but it declined in cluster analysis.

In conclusion, stepwise logistic regression and cluster analysis followed by logistic regression showed the best predictive power in this large and complex data base. The predictive power of stepwise logistic regression showed no major variation under different treatments of the missing values.

Bibliography

[1] Edwards, F.H. and Graeber, G.M., 'The theorem of Bayes as a clinical research tool', Surgery, 165, 127-9, (1987).

[2] Edwards, F.H., Albus, R.A., Zajtchuk, R., et al., 'Use of a Bayesian statistical model for risk assessment in coronary artery surgery', Ann Thorac Surg, 45, 437-440, (1988).

[3] Parsonnet, V., Dean, D., and Bernstein, A.D., 'A Method of Uniform Stratification of Risk for Evaluating the Results of Surgery in Acquired Adult Heart Disease', Circulation; 79 (suppl I): I-3-I-12, 1989.

[4] Brieman, L., Friedman, J.H., Olshen, R.A., Classification and Regression Tress, Wadsworth International Group, Belmont, California, 1984.

[5] Friedman, J.H., 'Multivariate Adaptive Regression Splines', Annal of Statistics, 19, 1-141,(1991).

[6] Hastie, T. and Tibshirani, R. Generalized Additive Models, Chapman and Hall, London, 1990.

[7] Harrell, F.E. Jr, Lee, K.L., Califf, R.M., et al., 'Regression Modelling Strategies for improved prognostic prediction'. Statistics in Medicine, 3, 143-152, (1984).

[8] Harrell, F.E. Jr, Lee, K.L., Matchar, D.B., et al., 'Regression Models for Prognostic Prediction : Advantages, Problems and Suggested Solutions', Cancer Treat Rep, 69, 1071:1077, (1985).

[9] Cook, E.F. and Goldman, L., 'Empiric comparison of multivariate analytic techniques: advantages and disadvantages of recursive partitioning analysis'. J Chronic Dis, 37, 721-731, (1984).

[10] Hanley, J.A. and McNeil, B.J., 'The Meaning and Use of the Area under a Receiver Operating Characteristic (ROC) Curve', Radiology, 143, 29-36, (1982).

[11] Grover, F.L., Hammermeister, K.E., Burchfiel, C., and Cardiac Surgeons of the Department of Veterans Affairs, 'Initial Report of the Veterans Administration Preoperative Risk Assessment Study for Cardiac Surgery', Ann Thorac Surg, 50, 12-28, (1990)

[12] Hsieh, F.Y., 'Sample size for logistic regression', Statistics in Medicine 8, 795-802, (1989).

[13] Lawless, J.F. and Singal, K., 'Efficient Screening of Nonnormal Regression Models', Biometrics, 34, 318-327, (1978).

[14] SAS Institute, Inc., 'SAS/STAT Guide for Personal Computers', Version 6 Edition, Cary, NC, (1985).

[15] Roberts, J.S. and Capalbo, G.M., 'A SAS macro for estimating missing values in multivariate data', Proceedings of the Twelfth Annual SAS User's Group International Conference, Dallas, Texas, 939-941, (1987).

Article 4.5

Center variability in risk of adjusted length of stay for very low birth weight infants in the Neocosur South American Network

Guillermo Marshall, Maria José Luque, Alvaro González, Ivonne D'Apremont, Gabriel Musante, José L. Tapia, et al.

Pontificia Universidad Catolica de Chile and
Neocosur Network

Abstract. To develop a prediction model for hospital length of stay (LOS) in very low birth weight (VLBW) infants and to compare this outcome among 20 centers within a neonatal network. Data from 7,599 infants with birth weights of 500-1,500 g born between the years 2001-2008 were prospectively collected. The Cox regression model was employed to develop two prediction models: an early model based upon variables present at birth, and a late one that adds relevant morbidities for the first 30 days of life. Median adjusted estimated LOS from birth was 59 days, 28 days after 30-day point of survival. There was a high correlation between models (r = 0.92). Expected to observed LOS varied widely among centers, even after correction for relevant morbidity after 30 days. Median observed LOS (range: 45-70 days), and postmenstrual age at discharge (range: 36.4-39.9 weeks) reflect high inter-center variability. A simple model, with factors present at birth, can predict a VLBW infant's LOS in a neonatal network. Significant variability in LOS was observed among neonatal intensive care units. We speculate that the results originate in differences in inter-center practices. [1]

Keywords: length of stay, discharge timing, hospital stay, very low birth weight infants.

1 Introduction

The remarkable improvement of very low birth weight (VLBW) infants survival observed over recent decades has been associated with an increased length of hospital stay (LOS). Both mortality and LOS are commonly used as quality of care measures for premature infants. In order to control for patient case-mix, comparisons among centers require procedures for risk adjustment. An unadjusted LOS has been described as a secondary outcome in several publications. Only few studies have focused on risk adjusted LOS as primary objective. Among these, single center studies (1; 2) are limited to predict LOS in that center, whereas multicenter studies (3; 4; 5) are able to predict and simultaneously compare LOS among centers. These studies have revealed significant variations in LOS between units (NICUs).

In the setting of a neonatal network, one of the available tools for quality improvement is to identify and compare factors that might influence variability between centers. Identifying the

[1] Marshall G, Luque MJ, Gonzalez A, D'Apremont I, Musante G, Tapia JL. Center variability in risk of adjusted length of stay for very low birth weight infants in the Neocosur South American Network. J Pediatr. 2012;88(6):524-30.

best-performing centers and examining their practices, may lead to the identification of potential interventions that can improve VLBW infant outcomes. Because risk factors vary across sites, statistical models should be used to adjust outcomes in order to compare center performances (3; 6). Despite documentation of variability in medical care during hospital stay, little is known about factors that might influence inter-NICU variation in LOS (4). Evaluation of medical practices for benchmarking purposes and quality of care comparisons require accurate and reliable risk models (7).

Prolonged LOS in VLBW infants carries several medical, psychosocial and economic problems (8). Each day of discharge delay accounts for a greater use of medical resources, NICU congestion, and consequently higher total costs. It also increases the risk of hospital-acquired morbidity, and may have an adverse effect on parenting by increasing the period of separation (9).

The NEOCOSUR network (https://neocosur.org/) is a voluntary non profit association of neonatal intensive care units (NICUs) from a group of South American countries (Argentina, Chile, Paraguay, Peru and Uruguay), whose primary objective is the continuous improvement of neonatal health. This network provides a continuous database that prospectively gathers information from all inborn VLBW (birth weight 500 to 1500 g) infants from the participating centers.

The purpose of this study was to develop prediction models for LOS among VLBW infants and compare this outcome in 20 participating centers from the Neocosur Neonatal Network.

2 Methods

We included all inborn infants with birth weight (BW) between 500 to 1500g admitted to the 20 NEOCOSUR Network centers from January 1, 2001 to December 31, 2008. This Neonatal Network includes Neonatal Intensive Care Units (NICUs) from 5 of the most southern countries in South America: Argentina, Chile, Paraguay, Peru and Uruguay.

Only inborn infants who were admitted and completed their stay (either by discharge home or death) at each NICU were included in the analyses. Infants who were transferred to other NICU after admission were not included. Demographic and clinical information and outcome data were prospectively and routinely collected at the NEOCOSUR Network centers using predefined diagnostic criteria and online data entry. In order to evaluate factors that may influence and predict LOS, we developed two models:

1. An early model including all cases, considered only variables present at birth (before NICU admission) such as: birth weight, postmenstrual age (PMA), 1 minute Apgar score, gender, presence of multiple birth, antenatal steroids use, presence of congenital malformations and prenatal care.
2. In order to identify further factors affecting LOS, we developed a late model which considered additional relevant in-hospital morbidities or clinical events occurring during the first 30 days of hospitalization: respiratory distress syndrome, mechanical ventilation, bronchopulmonary dysplasia (BPD), severe (Grade III or IV) intraventricular hemorrhage (IVH), early and late onset sepsis, patent ductus arteriosus (PDA), necrotizing enterocolitis (NEC). This model included only those infants whose length of stay was greater than 30 days. BPD was defined as oxygen therapy for 28 days or more after birth. The diagnosis of late onset sepsis was confirmed by the isolation of the organism in blood or cerebrospinal fluid after 72 hours of life. Patent ductus arteriosus was diagnosed clinically and whenever possible confirmed by

echocardiography. The diagnosis of IVH was made by cranial ultrasonogram or by autopsy and was classified according to Papile et al (10). Necrotizing enterocolitis was confirmed by radiological (pneumatosis and/or perforation), surgical or autopsy findings.

Employing these factors we developed a prediction score for both models. Initially, univariate associations between infant related variables and LOS were performed using a simple Cox's regression model. Mortality was included, and LOS were censored at the time of death of the individual case. Then, a stepwise multiple Cox's regression model was used to select the subset of variables that were independently associated with LOS (11). A significance level of 5% was used to include each variable in the model.

With the development of the final 2 models, we were able to estimate the risk-adjusted LOS by calculating the Cox with regression coefficients and applying them to Kaplan Meier type estimators of the stay curve (11).

Overall network and center specific LOS stay functions were calculated adjusting by each infant's relative risk. The resulting curves for each center were directly comparable since they were corrected by differences in patient mix. In order to compare center performances we calculated the median LOS obtained from each risk adjusted stay functions (Observed - Expected LOS). PMA at discharge among survivors was also analyzed by risk quartiles (based on the early model) and compared among centers. The R software (R Foundation for Statistical Computing, Vienna, Austria) was used for all statistical calculations (12).

3 Results

Data from 7,599 inborn infants were analyzed. BW was 1,101 ± 271 g (mean±SD) and gestational age (GA) was 29.2 ± 2.9 weeks (mean±SD). Female gender rate was 48.9% (44.9-57.9% range). Rate of antenatal steroids was 74.7% (42.1-91.1% range). Multiple gestations were present in an 18.4% (10.9-41.1% range). Rate of cesarean section was 52.1% (28.6-68.2% range). The percentage of Apgar score ≤ 3 at 1 minute was 19.5% (7.9-33.3% range) and at 5 minutes was 3.2% (0-6.7%). The total mortality rate was 24% with a range from 10 to 47.7%. The total incidence of BPD at 28 days was 23.4% (5.3-38.7% range).

BW was the most important factor, emphasizing that each additional 100 g increases the likelihood of being discharged from the hospital by 22.4%. The second factor (in order of importance) was PMA, indicating that every additional week in PMA increases the likelihood by 11.5% of being discharged. These significant factors were the same found while developing the Neocosur score for predicting mortality in VLBW infants (7), with only "absence of life" threatening the position of "congenital malformations." Figure 1A shows the LOS curve from birth for the overall network. Length of stay for very low birth weight infants - Marshall G et al.

Table 1 also shows the factors selected for the second regression model used to describe LOS after 30 days of survival. As expected, greater BW and PMA were associated with shorter LOS (or an increased possibility of being discharged), whereas the presence of BPD, NEC, severe IVH, sepsis, and PDA were associated with significantly longer LOS (or decreased chance of being discharged). BPD was the most significant factor decreasing the likelihood of being discharged by 46.2%. Figure 1B illustrates the LOS curve among infants who remained at the NICUs for more than 30 days within the overall network.

Table 1. Factors present at birth selected by a Stepwise Cox's Regression Model to estimate LOS

			Relative Risk	
Variables	Coefficient*	SE	Estimate*	95% CI
Factors present at birth				
Birth Weight per 100g	0.2024	0.0073	1.224	1.21 - 1.24
Gestational Age per week	0.1093	0.0068	1.115	1.10 - 1.13
1-minute Apgar	0.0381	0.0067	1.039	1.03 - 1.05
Antenatal steroid use	0.0760	0.0157	1.079	1.05 - 1.11
Female gender	0.1009	0.0269	1.106	1.05 - 1.17
Factors after 30 days of survival				
Birth Weight per 100g	0.166	0.0075	1.18	1.16 - 1.20
BDP	-0.6205	0.0367	0.54	0.50 - 0.58
Gestational Age per week	0.0482	0.0073	1.05	1.03 - 1.06
Late Onset Sepsis	-0.2377	0.0364	0.79	0.73 - 0.85
PDA	-0.1913	0.0328	0.826	0.77 - 0.88
NEC	-0.2821	0.0494	0.75	0.68 - 0.83
IVH Grade III and IV	-0.2818	0.0625	0.75	0.67 - 0.85

* In a Cox's regression model a positive coefficient and a $RR < 1$ are associated to a higher risk of being discharged or lower LOS.

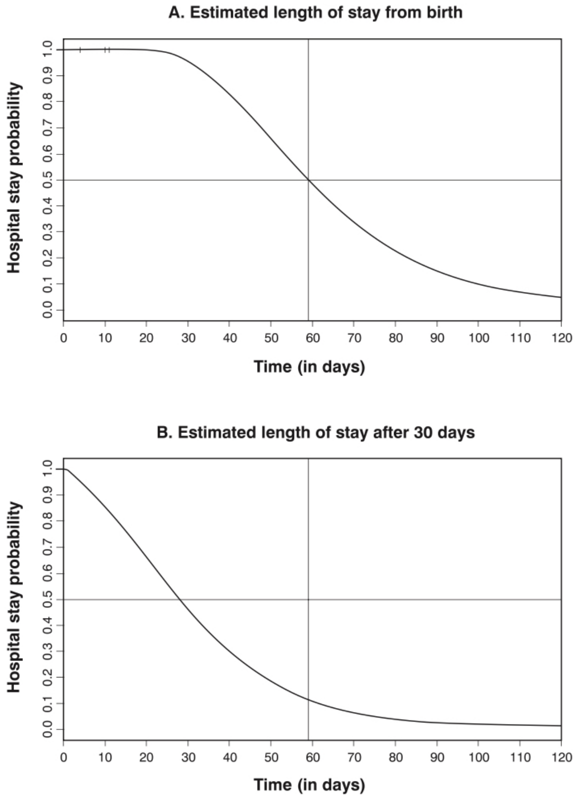
High variability in risk-adjusted median LOS was observed among the 20 Neocosur NICUs via calculations with both models. The overall median and range of LOS was 59 (45- 70) days after birth, and 28 (18-38) days after the first 30 days of survival. Eight centers within the network were observed to have higher LOS than the median, 11 centers were below it, and one center had the same median as the median LOS for the entire network. This variability was still high when analyzing the remaining LOS after 30 days of survival. Figure 2 (2A and 2B) illustrate such high variability, showing how each individual center compares to the overall Neocosur network.

Most of the centers had important differences between the observed and expected LOS, some of them discharging their patients earlier (negative Observed - Expected LOS) and others exhibiting more prolonged hospital stays (positive Observed - Expected LOS), when compared to the overall median risk-adjusted LOS of the network. When risk-adjusted LOS at birth and after 30 days were compared among the centers, a high correlation of r = 0.92 was found, showing that both are consistent indicators.

When we analyzed the median PMA at discharge, thus categorizing infants by risk quartiles using the LOS Neocosur score at birth, we also found significant variability among centers (Table 2).

4 Discussion

We developed two prediction models for risk-adjusted LOS in a population of VLBW infants. The early model included only predictors present at birth while the late one included relevant



A. Estimated length of stay from birth

B. Estimated length of stay after 30 days

Fig. 1. Hospital stay probability (stay curve) in the Neocosur network from birth (Figure 1A) and after 30 days (Figure 1B) for different time points (in days). In each point of time, these probabilities represent the expected proportion of newborn infants that remain hospitalized. The vertical line marks where 50% (median) of the infants remain hospitalized

morbidity during the first 30 days of life.

In both models the most important variable for predicting LOS was BW, which is also the principal factor for predicting in-hospital mortality (7). The correlation between LOS and BW has already been described (4; 13; 14). The coincidence between the main factors for predicting both LOS and mortality reinforces the finding that infants delivered with lower BWs do not only have greater risks of mortality, but also face longer LOS if they survive. This is likely to be due to immaturity and also to a higher incidence of medical complications in this group of patients. In fact, in the late model the most important factors for LOS prediction, other than BW, were BPD and sepsis. It is of interest that the use of prenatal steroids was also a factor associated with a shorter LOS, giving weight to the other beneficial effects associated with this therapy that have been widely reported in the literature (15).

When we compared the observed versus expected LOS at each NICU, we found high variability among centers and important differences in LOS between those NICUs for infants with

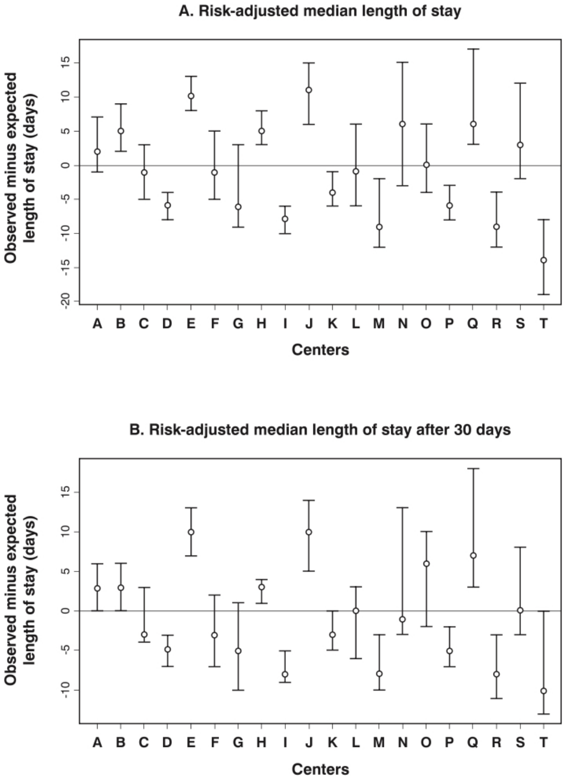


Fig. 2. Observed minus expected hospital length of stay (in days) and 95% confidence intervals for the 20 neonatal intensive care units calculated from birth (Figure 2A) and after 30 days (Figure 2B). The horizontal line shows where the adjusted hospital length of stay of an individual center is equal to the entire Neocosur network

similar risks. This variability in LOS among centers was not significantly reduced after adjustments based on the late model at a later time during hospital stay. Our interpretation of this finding is that other factors, such as center practices, may better explain this variability. Risk-adjusted LOS also varied significantly between NICUs with regard to PMA at discharge. The interval between the earliest and latest discharging NICU was 3 to 4 weeks of PMA in all risk quartiles. Additionally, when risk quartiles were compared, we observed a consistent difference in PMA between the lowest and highest risk groups. As expected, high-risk infants faced longer LOS and, consequently, were discharged at higher PMAs. However, in intermediate risk quartiles, this pattern was not always consistent at all the centers, suggesting inter-NICU variability in infants with the same risks.

Factors that might influence inter-NICU variability are differences in medical care during hospitalization, different discharge policies, in-hospital morbidities, population differences, availability of home care, and community support. Eichenwald et al.4 studied LOS in a homogeneous healthy population of premature infants delivered at 30.0 to 34.6 weeks of gestation. PMA at discharge varied between 35.2 to 36.5 weeks. The authors concluded that inter-NICU variation in

Table 2. Postmenstrual age at discharge by risk quartiles (based on the Neocosur Score, 1=lower risk, 4=higher risk) and the total of each center.

	Risk Quartiles				
NICU	1	2	3	4	Total
A	37.9	37.4	38.9	31	28-34
B	37.3	36.9	38.3	31	28-34
C	38.9	38.0	38.0	25	24-31
D	38.0	36.7	38.1	23	21-25
E	39.3	38.6	39.3	38	35-41
F	36.7	37.0	38.1	25	21-30
G	38.1	37.6	39.7	23	18-29
H	39.0	38.4	40.1	31	29-32
I	36.9	36.4	36.6	20	19-23
J	38.7	39.0	38.9	38	33-42
K	37.7	36.6	37.7	25	23-28
L	37.4	36.6	37.9	25	22-31
M	37.0	36.0	37.1	25	18-25
N	39.1	38.1	39.1	25	25-41
O	36.3	36.9	38.3	25	26-38
P	37.4	37.1	37.7	25	21-26
Q	40.1	39.3	39.9	25	31-46
R	36.9	37.4	36.7	25	17-25
S	37.9	37.9	38.6	25	25-36
T	36.1	36.2	37.4	25	15-28
Network	37.9	37.3	38.3	25	27-29

recorded maturational milestones (mature feeding behavior, cessation of apnea, and bradycardia events) was the most significant influence on LOS. They also suggest that variation in care practices, rather than differences in clinical characteristics, contributed to differences in discharge timing between hospitals (4). In another study, Cotten et al. (16) analyzed center-independent factors associated with prolonged hospital stays (PHS) in extremely premature infants, and concluded that chronic lung disease, surgical NEC, and late onset sepsis are variables that contribute to PHS. Similarly, in a previous publication from Neocosur,17 we also found longer LOS in VLBW infants who developed BPD compared to those who did not.

A recent NICHD Network study compared several models for predicting time of hospital discharge for extremely preterm infants (less than 27 weeks of gestational age), and concluded that prediction of early or late discharge is poor when only perinatal factors are considered. However, predictability can substantially improve with knowledge of later-occurring morbidities (18). In contrast, the present study shows that the adjusted early prediction model is strongly correlated (r = 0.92) with the late corrected model at 30 days of life. Although the late model yielded more accurate predictions, our data show that center variability in LOS remains similar even after adjusting the model with a selection of major morbidities (BPD, IVH III-IV, NEC, PDA, and late onset sepsis) developed during the first 30 days of hospital stay. One explanation for this high correlation is that risk factors present at birth may also determine the appearance of later in-hospital complications. However, the persistence of LOS differences between centers, regardless of corrections by risks or major morbidities, suggests that local center factors play a role in determining final LOS. We could speculate that differences in clinical management among centers may have constituted the principal factor influencing LOS variability in our study. We must also consider that centers in the NIH network are quite homogeneous, while Neocosur centers differ greatly in terms of resources, size, case-mix, within other potentially relevant factors. Finally, we should consider that discharge timing of premature infants is a complex process influenced not only by medical factors, but also by nonmedical issues such as primary healthcare and organizational delays, discharge planning delays, as well as family circumstances, among other factors..

Another limitation in this study was the fact that information regarding the various factors that can delay discharge, such as duration of apnea or feeding problems, was not available. Also, as mentioned previously, the centers in this network have important variations in other major outcomes. The published experience in all networks shows large outcome variability. Studying these institutional differences is beyond the scope of this study.

This, however, is a true representation of the reality of our region. This study benefits from the inclusion of a large multicenter population, in contrast to several studies that referred to one center only, and this allows us to perform a benchmarking analysis.

We conclude that LOS for VLBW infants can be successfully predicted in a neonatal network by using two prediction models: one at birth and another at 30 days of life. The early model has the advantage of constituting a simple infant score with factors observed at birth. Thus, early estimations of LOS can be useful for families and medical care providers. Our models also enable us to predict and compare LOS among centers. The results reveal significant inter-NICU LOS differences, entailing an important economic impact, which is highly relevant in a region with limited resources. Center comparisons may also contribute to strategic planning that might safely reduce LOS for VLBW infants in some centers thus decreasing hospitalization costs and risks associated for prolonged hospitalization in this vulnerable BW group.

Acknowledgements

We thank all the Neocosur centers that participated in this study. The present study included the following collaborators from the Neocosur Network:

Argentina: Guillermo Colantonio, Gabriel Musante, Luis Prudent, Liliana Rochinotti, Ines Galindez, Mariana Sorgetti, Lorena Soler (Clínica y Maternidad Suizo Argentina, Buenos Aires); Isabel Kurlat, Oscar Di Siervi, Adriana Escarate (Hospital de Clinicas José de San Martin, Buenos Aires); Gonzalo Mariani, Jose María Ceriani, Silvia Fernandez, Carlos Fustiñana (Hospital Italiano, Buenos Aires); Jorge Tavosnaska, Liliana Roldan, Hector Sexer, Elizabeth Lombardo (Hospital Juan Fernandez, Buenos Aires); Gabriela Torres, Daniel Agost, Augusto Fischetti, Monica Rinaldi (Hospital Lagomaggiore, Mendoza); Carlos Grandi, Claudio Solana, Javier Meritano, Miguel Larguia (Maternidad Sarda, Buenos Aires), Marcelo Decaro, Lionel Cracco, Gustavo Bassi, Noemi Jacobi, Andrea Brum, Nestor Vain (Sanatorio de la Trinidad, Buenos Aires); Adriana Aguilar, Miriam Guerrero, Edgardo Szyld, Alcira Escandar (Hospital Dr. Diego Paroissien, Buenos Aires); Horacio Roge, Mario Marsano, Elisa Fehlmann, Jorge Rós (Hospital Español de Mendoza, Mendoza);

Chile: Jorge Fabres, Alberto Estay, Alvaro Gonzalez, Sandra Vignes, Mariela Quezada, Jose L. Tapia, Soledad Urzua (Hospital Clinico Universidad Catolica de Chile, Santiago); Rodrigo Ramírez, Maria Eugenia Hubner, Jaime Burgos, Jorge Catalan (Hospital Clinico Universidad de Chile, Santiago); Lilia Campos, Aldo Bancalari, Lilian Cifuentes, Jorge Leon, Eduardo Broitman, Roxana Aguilar (Hospital Guillermo Grant, Concepcion); Jane Standen, Marisol Escobar, Alejandra Nuñez (Hospital Gustavo Fricke, Viña del Mar); Agustina González, Ana Luisa Candia, Lorena Tapia, Giovanna Loguercio, Claudia Avila (Hospital San Jose, Santiago); Claudia Toro, Patricia Mena, Angelica Alegria, Adolfo Llanos (Hospital Dr. Sótero del Rio, Santiago); Veronica Peña, Marianne Bachler, Patricia Duarte (Hospital San Borja Arriaran, Santiago); Ivonne D'Apremont, Guillermo Marshall, Sandra Vignes, Mariela Quezada, Luis Villarroel, Angelica Dominguez (Unidad Base de Datos, Pontificia Universidad Católica, Santiago);

Paraguay: Jose Lacarruba, Elizabeth Cespedes, Ramon Mir, Elvira Mendieta, Larissa Genes, Carlos Caballero (Departamento de Hospital de Clinicas de Asuncion, Asuncion);

Perú: Jaime Zegarra, Veronica Webb, Fabiola Rivera, Marilu Rospigliosi, Silvia Febres, Enrique Bambaren (Hospital Cayetano Heredia, Lima); Rosa Unjan, Walter Cabrera, Raul Llanos, Anne Castañeda, Oscar Chumbes, Roberto Rivera (Hospital Guillermo Almenara, Lima);

Uruguay: Ruben Panizza, Sandra Gugliucci, Silvia Fernandez, Eduardo Mayans, Alicia Prieto, Cristina Hernandez (Facultad de Medicina Servicio de Recien Nacidos, Montevideo).

Bibliography

[1] Zernikow B, Holtmannspotter K, Michel E, Hornschuh F, Groote K, Hennecke H. Predicting length-of-stay in preterm neonates. Eur J Pediatr 1999; 158: 59-62.

[2] Powell P, Powell C, Hollis S, Robinson MJ. When will my baby go home? Arch Dis Child 1992; 67: 1214-1216.

[3] Berry MA, Shah PS, Brouillette RT, Hellmann J. Predictors of mortality and length of stay for neonates admitted to children´s hospital neonatal intensive care units. J Perinatol 2008; 28(4): 297-302.

[4] Eichenwald EC, Blackwell M, Lloyd JS, Tran T, Wilker RE, Richardson DK. Inter-neonatal intensive care unit variation in discharge timing: influence of apnea and feeding management. Pediatrics 2001; 108(4): 928-33.

[5] Merenstein D, Egleston B, Diener-West M. Lengths of stay and costs associated with children´s hospitals. Pediatrics 2005; 115(4): 839-44.

[6] Richardson DK, Tarnow-Mordi WO, Escobar GJ. Neonatal risk scoring systems. Can they predict mortality and morbidity? Clin Perinatol 1998; 25: 591-611.

[7] Marshall G, Tapia JL, D´Apremont I, Grandi C, Barros C, Alegria A et al. A new score for predicting neonatal very low birth weight mortality risk in the NEOCOSUR South American Network. J Perinatol 2005; 25: 577-582.

[8] Casiro OG, McKenzie ME, McFadyen L, Shapiro C, Seshia MM, MacDonald N et al. Earlier discharge with community-based intervention for low birth weight infants: a randomized trial. Pediatrics 1993; 92(1): 128-34.

[9] American Academy of Pediatrics, Committee on Fetus and Newborn. Hospital discharge of the high-risk neonate–proposed guidelines. Pediatrics 1998; 102(2 Pt 1): 411-7.

[10] Papile L, Bursten J. Incidence and evolution of subependimal and intraventricular hemorrhage. A study of infants with birth weights less than 1500g. J Pediatr 1978; 92: 529-534.

[11] Collett D. (1994) Modeling Survival Data in Medical Research, London: Chapman & Hall.

[12] Development Core Team (2003). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL, http://www.R-project.org .

[13] Rawlings JS, Scott JS. Postconceptional age of surviving preterm low birth-weight infants at hospital discharge. Arch Pediatr Adolesc Med 1996; 150: 260-262.

[14] Bannwart Dde C, Rebello CM, Sadeck LS, Pontes MD, Ramos JL, Leone CR. Prediction of length of hospital stay in neonatal units for very low birth weight infants. J Perinatol 1999; 19(2): 92-6.

[15] Crowley P. Prophylactic corticosteroids for preterm birth. Cochrane Database Syst Rev 2006; 18(33): CD000065.

[16] Cotten CM, Oh W, McDonald S, Carlo W, Fanaroff AA, Duara S et al. Prolonged hospital stay for extremely premature infants: risk factors, center differences, and the impact of mortality on selecting a best-performing center. J Perinatol 2005; 25(10): 650-5.

[17] Tapia JL, Agost D, Alegria A, Standen J, Escobar M, Grandi C et al. Bronchopulmonary dysplasia: incidence, risk factors and resource utilization in a population of South American very low birth weight infants. J Pediatr (Rio J) 2006; 82(1): 2-3.

[18] Hintz SR, Bann CM, Ambalavanan N, Cotten CM, Das A, Higgins RD. Predicting time to hospital discharge for extremely preterm infants. Pediatrics 2010; 125(1): 146-54.