Multivariate Modelling of a Monotone Disease Process in the Presence of Misclassification

María José García-Zattera\textsuperscript{1,2,\dagger}, Alejandro Jara\textsuperscript{1}, Emmanuel Lesaffre\textsuperscript{2,3} and Guillermo Marshall\textsuperscript{1}

\textsuperscript{1} Department of Statistics, Pontificia Universidad Católica de Chile, Casilla 306, 22 Santiago, Chile
\textsuperscript{2} L-BioStat, Katholieke Universiteit Leuven, Kapucijnenvoer 35, Block D, Bus 7001, B-3000 Leuven, Belgium
\textsuperscript{3} Department of Biostatistics, Erasmus Medical Center, Dr Molewaterplein 50, 3015 GE Rotterdam, the Netherlands
\dagger Email: mjgarcia@uc.cl

Abstract: Motivated by a longitudinal oral health study, the Signal-Tandmobiel\textsuperscript{R}, we propose a multivariate binary inhomogeneous Markov model in which correlated true response variables are subject to an unconstrained misclassification process and follow a monotone behavior. Conditionally specified logistic regression models are used to relate the multivariate baseline distribution and the elements of the transition matrices of the unobserved process to the available covariates. The misclassification model considers the existence of different classifiers for each subject across time.

Keywords: Multivariate Binary Data; Monotone Processes; Misclassification; Incidences Estimation; Hidden Markov Model.

1 Introduction

The motivation for this work comes from data gathered in a longitudinal oral health study conducted in Flanders (Belgium) between 1996 and 2001, the Signal-Tandmobiel\textsuperscript{R} (ST) study (Vanobbergen et al., 2000). For this project, 4,468 children were examined on a yearly basis during their primary school time (between 7 and 12 years of age) by one of sixteen dental examiners. The purpose of the investigation is to examine the association of several potential risk factors with the prevalence and incidence of caries experience (CE), which is defined as a binary variable indicating whether a tooth is decayed at $d_3$ level, missing or filled due to caries. The assessment of risk factors under this setting involves the analysis of a misclassified multivariate monotone binary process since: i) CE, as previously defined, is a progressive or monotone disease because teeth cannot alternate between the presence or absence of CE once CE is observed over time, ii) events on
teeth of the same child are dependent, and iii) the examiners scoring may not perfectly reflect the tooth's true condition and the presence of CE can be miss-diagnosed.

In this work we extend recent developments on univariate hidden binary monotone Markov models (García-Zattera et al., 2010) to provide a framework for the analysis of multivariate hidden monotone process as a function of covariates. The rest of the manuscript is organized as follows. In Section 2 we introduce the proposed model. Section 3 presents the analysis of our motivating problem. A final discussion section concludes the manuscript.

2 The model

Suppose that \( J \) teeth are examined on subject \( i, i = 1, \ldots, I \), at time points \( t(1), k = 1, \ldots, K \). Let \( Y(i,k) \) be the true unobserved binary response for tooth \( j \) of subject \( i \) at time \( t(i,k) \) and denote the \( J \)-dimensional vector of true responses for all teeth of subject \( i \) at time \( t(i,k) \) by \( Y(i,k) = (Y(i,1), \ldots, Y(i,J))^T \). Let \( x(i,j,k) \) be a \( p \)-dimensional vector of covariates for subject \( i \) and tooth \( j \) at examination \( k \). We assume that the vectors \( Y(i,k) \) follow a monotone inhomogeneous first-order Markov process. In order to relate the covariates to the initial distribution and to the elements of the transition matrices in a population-average manner and taking into account the association among the responses of the same subject, conditionally specified logistic regression models (Joe and Liu, 1996; García-Zattera et al., 2007) are used. Specifically, we assume that

\[
\Pr(Y(i,1) = y \mid X(i,1)) \propto \exp \left\{ \sum_{j=1}^{J} (x(i,j,1)' \beta_P^P) y_{j} + \sum_{1 \leq j < l \leq J} \gamma_P y_{j} y_{l} \right\},
\]

and, for \( k = 2, \ldots, K \),

\[
\Pr(Y(i,k) = y^1 \mid Y(i,k-1) = y^0, Z(i,k)) \propto \exp \left\{ \sum_{j=1}^{J} (z(i,j,k)' \beta_P^I) y^1_{j} (1 - y^0_{j}) + \sum_{1 \leq j < l \leq J} \gamma_I y^1_{j} y^1_{l} (1 - y^0_{j} y^0_{l}) + \sum_{j=1}^{J} \sum_{l \neq j} \alpha_I y^0_{j} y^1_{l} (1 - y^0_{l} y^0_{l}) \right\}, \tag{1}
\]

where \( y \in \{0,1\}^J, y^1 \in B(y^0) \subset \{0,1\}^J \), with \( B(y^0) \) being an admissible set, \( z(i,j,k) = (x(i,j,k-1)' , t(i,k) - t(i,k-1)) \). \( \beta_P^P \) and \( \beta_I^I \) are vectors of conditionally specified logistic regression models associated to the initial distribution and transition matrices, respectively, \( \gamma_P^P \) and \( \gamma_I^I \) are vectors of within-time conditional log-odds ratio parameters for the initial distribution and transition matrices, respectively, and \( \alpha_I^I \) is a vector of across-time
conditional log-odds ratios. We show that expression (1) defines a proper probability model for each row in the Markovian transition matrices.

We assume that the response variables $Y_{(i,j,k)}$ are prone to misclassification. Let $Y^*_{(i,j,k)}$ be the corrupted observed binary response for tooth $j$ of subject $i$ at time $t_{(i,k)}$. Here we suppose that the scoring is performed by $Q$ examiners. Denote by $\xi_{i,k} \in \{1, \ldots, Q\}$ the indicator variable of examiner that scores all teeth of subject $i$ at time $t_{(i,k)}$.

In an initial version of the model ($M_1$) we assume that the scoring behavior of the examiners is the same across the study and teeth. Let $\tau_{e,11}$ and $\tau_{e,00}$ be the sensitivity and specificity, respectively, for examiner $e$, $e = 1, \ldots, Q$.

The misclassification model assumes that $\Pr(Y^*_{(i,j,k)} = 1 \mid Y_{(i,j,k)} = 1) = \tau_{\xi_{i,k},11}$ and $\Pr(Y^*_{(i,j,k)} = 0 \mid Y_{(i,j,k)} = 0) = \tau_{\xi_{i,k},00}$. In a second version of the model ($M_2$), we assume that the scoring behavior of the examiner is tooth-specific, but the same across the study; i.e., here $\tau_{e,j,11}$ and $\tau_{e,j,00}$, $j = 1, \ldots, J$, denote the tooth-specific sensitivity and specificity, respectively, for examiner $e$. By using simulated data, we show that, under the setting of the motivating problem, the restriction $\tau_{e,00} + \tau_{e,11} > 1$ for $M_1$ and $\tau_{e,j,00} + \tau_{e,j,11} > 1$ for $M_2$, along with restrictions on the rank of the design matrices, are sufficient conditions in order to ensure parameters identifiability.

3 The analysis of the Signal-Tandmobiel® data

We implement Bayesian versions of the models and assume independent $N(0, 10^3)$ priors for the coordinates in $\beta^P, \beta^I, \gamma^P, \gamma^I$ and $\alpha^I$. For the misclassification parameters, independent uniform distributions are assumed, under the restriction $\tau_{e,00} + \tau_{e,11} > 1$ for $M_1$ and $\tau_{e,j,00} + \tau_{e,j,11} > 1$ for $M_2$. We fit the models to the four permanent first molars, teeth 16, 26 on the maxilla (upper quadrants), and teeth 36 and 46 on the mandible (lower quadrants). We evaluated the effect of the age at start of brushing (in years), the number of between-meal snacks (two or less than two a day vs more than two a day), the geographical location, and the age (in years) on the prevalence and incidences of CE. Model comparison was performed using the pseudo Bayes factor (PsBF) (Geisser and Eddy, 1979).

The $2\times \log_{10}$PsBF for $M_1$ versus $M_2$ was 40.6, suggesting no evidence for the hypothesis of tooth-specific misclassification parameters for each examiner. Therefore, we only report the results arising from $M_1$. Table 1 shows the posterior mean and 95% highest posterior density (95% HPD) credible intervals for the conditionally specified logistic regression coefficients. The results suggest that the older the child, the higher the prevalence of CE, and that the later the child starts brushing or the higher the number of between-meal snacks, the higher the incidence of CE.
TABLE 1. Posterior mean and 95% highest posterior density (95% HPD) credible intervals, for the conditionally specified logistic regression coefficients associated to the prevalence and incidences for CE in permanent first molars.

<table>
<thead>
<tr>
<th></th>
<th>Prevalence</th>
<th>Incidences</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Posterior Mean</td>
<td>95%HPD</td>
</tr>
<tr>
<td>Intercept T16</td>
<td>-6.16 (-8.78；-3.54)</td>
<td>-4.62 (-5.53；-3.59)</td>
</tr>
<tr>
<td>Intercept T26</td>
<td>-5.62 (-8.20；-3.13)</td>
<td>-4.33 (-5.25；-3.37)</td>
</tr>
<tr>
<td>Intercept T36</td>
<td>-5.75 (-8.36；-3.26)</td>
<td>-4.65 (-5.62；-3.63)</td>
</tr>
<tr>
<td>Intercept T46</td>
<td>-5.58 (-8.07；-3.02)</td>
<td>-4.21 (-5.12；-3.24)</td>
</tr>
<tr>
<td>Age Start Brushing</td>
<td>0.09 (-0.02；0.20)</td>
<td>0.09 (0.04；0.13)</td>
</tr>
<tr>
<td>Age</td>
<td>0.35 (0.00；0.69)</td>
<td>-0.03 (-0.08；0.01)</td>
</tr>
<tr>
<td>Meals</td>
<td>0.15 (-0.11；0.41)</td>
<td>0.15 (0.05；0.27)</td>
</tr>
<tr>
<td>x-ordinate</td>
<td>0.08 (-0.24；0.25)</td>
<td>0.07 (-0.04；0.17)</td>
</tr>
<tr>
<td>y-ordinate</td>
<td>0.00 (-1.43；0.18)</td>
<td>-0.16 (-0.50；0.17)</td>
</tr>
<tr>
<td>Years Between Exam.</td>
<td>0.39 (0.00；0.82)</td>
<td>0.39 (-0.08；0.82)</td>
</tr>
</tbody>
</table>

Posterior mean and 95% HPD credible intervals for the association parameters are shown in Table 2. The posterior inferences for the within-time conditional log-odds suggest a high positive association in the presence of CE between symmetrically opponent molars and right vertically opponent molars (maxilla versus mandible) at the age of 7. At this age, a non-significant conditional association was found between diagonally opponent teeth. High positive within-time conditional associations were found between symmetrically, right vertically opponent molars and diagonally opponent teeth as the process evolves.

The posterior inference for the across-time odds ratio parameters suggest significant and negative associations between symmetrically and diagonally opponent molars. These results suggest that the probability of developing caries on a tooth is higher when a symmetrically or diagonally opponent molar is affected at the same time but sound at the previous examination, than when it was affected in the previous examination. This result can be explained by the fact that once a tooth is affected by caries, it is probably treated and the infection is no longer spreading in the next examination.

Finally, Figure 1 shows the posterior mean and 95% HPD credible intervals for the sensitivity and specificity for each examiner. The results suggest a greater variability in the sensitivity than in the specificity estimates, which can be explained by the low prevalence and incidences of CE. All examiners showed a sensitive greater than 0.7, with the exception of one examiner who showed a rather poor scoring behavior. The latter result is explained by the lower information available for this examiner. In fact, this examiner was only involved in the first two years of the ST study. The posterior mean for the specificity parameters were higher than 0.96 for all examiners.
TABLE 2. Posterior mean and 95% highest posterior density (95% HPD) credible intervals of conditional odds ratios for CE in permanent first molars.

<table>
<thead>
<tr>
<th></th>
<th>Prevalence</th>
<th>Incidences</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>16;26</td>
<td>16;26</td>
</tr>
<tr>
<td>Within</td>
<td>3.95 (2.84;5.13)</td>
<td>3.86 (3.18;4.55)</td>
</tr>
<tr>
<td>Time</td>
<td>0.59 (-1.74;2.92)</td>
<td>2.35 (1.25;3.38)</td>
</tr>
<tr>
<td>Association</td>
<td>2.30 (0.27;4.07)</td>
<td>1.11 (0.19;2.09)</td>
</tr>
<tr>
<td>Parameters</td>
<td>1.57 (-1.05;3.72)</td>
<td>0.63 (-1.94;0.78)</td>
</tr>
<tr>
<td></td>
<td>0.21 (-2.52;2.34)</td>
<td>2.11 (1.15;2.99)</td>
</tr>
<tr>
<td></td>
<td>2.63 (1.36;3.91)</td>
<td>3.71 (3.03;4.39)</td>
</tr>
</tbody>
</table>

FIGURE 1. Posterior mean and 95% highest posterior density credible intervals for examiner’s sensitivity (panel a) and specificity (panel b).
4 Concluding Remarks

We have proposed a multivariate hidden Markov model for monotone binary processes, which describes the relationships between covariates and the prevalence and incidences in a population-average fashion, and where different classifiers are present. The association between the responses, is taken into account by within- and across-time conditional odds ratios. An advantage of the proposed model is that the parameters can be estimated without external information on the misclassification parameters. Results from simulation studies suggest that even under the use of uniform priors on the misclassification parameters, unbiased and precise estimates of the parameters can be obtained.

References


