Computational Statistics and Data Analysis 53 (2009) 2724–2732

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

# Bayesian analysis of two dependent 2 × 2 contingency tables☆

Anastasia G. Eleftheraki [a], Maria Kateri [a,*], Ioannis Ntzoufras [b]

[a] Department of Statistics and Insurance Science, University of Piraeus, Piraeus, Greece
[b] Department of Statistics, Athens University of Economics and Business, Athens, Greece

ARTICLE INFO

*Article history:*
Received 8 May 2008
Received in revised form 21 January 2009
Accepted 21 January 2009
Available online 29 January 2009

ABSTRACT

Bayesian analysis of correlated binary data when individual information is not available is considered. In particular, a binary outcome is measured on the same subjects of two independent groups at two separate occasions (usually time points). The groups are formulated through a binary exposure or a prognostic factor. Interest lies in estimating the association between exposure and outcome over time. Standard methods for this purpose apply on the individual item responses and are insufficient in case these are missing. Moreover it is assumed that the only available information is the marginal 2 × 2 cross-tabulations between the grouping variable and the response for each occasion. Assuming independent binomial distributions for the two groups, the success probabilities for each occasion as well as the associations between exposure and outcome, based on the corresponding odds ratios, are estimated. In order to deal with the missing information of each item's response and to estimate the corresponding transition probabilities, a Bayesian procedure is adopted.



## 1. Introduction

The evaluation of the influence of an exposure or a prognostic factor on a response variable, is often the issue in clinical applications. In the characteristic case where both, exposure and response, are binary, the data are summarized in a 2 × 2 table. The measures used for inference on the response of the two groups formulated by the presence or not of the prognostic factor are the relative risk and the odds ratio.

If the binary response is measured repeatedly over time on the same subjects, more advanced tools are applied for the analysis. In the framework of longitudinal studies many models have been developed to analyze repeated measurements. Most of them model the marginal expectation of the binary responses using the generalized estimating equations (GEE) methodology (Liang and Zeger, 1986). For example, Fitzmaurice and Laird (1993) modeled time-dependent multivariate binary data by a marginal logistic model. Bayesian models using MCMC algorithms for the analysis of longitudinal data are recommended by Chib and Carlin (1999) while Tan et al. (2006) avoided the convergence problem associated with the MCMC approach by developing a non-iterative sampling method (the Inverse Bayes formulae sampler) for making Bayesian inference in hierarchical models for repeated binary data. In a different framework, Agresti and Klingenberg (2005) proposed a multivariate test comparing the binomial probabilities of two groups in a safety study where each subject was examined in some adverse events. Recently, Jara et al. (2007) incorporated latent variables and applied a semiparametric Bayesian approach in order to model multivariate correlated binary data assuming non-homogeneous structure across the subjects.

☆ The program code accompanied by a short help file are available at http://stat-athens.aueb.gr/~jbn/papers/paper22.htm.

* Corresponding address: Department of Statistics and Insurance Science, University of Piraeus, 80, Karaoli & Dimitriou Str., GR 185 34Piraeus, Greece. Tel.: +30 210 4142467; fax: +30 210 4142340.
*E-mail address:* mkateri@unipi.gr (M. Kateri).

0167-9473/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2009.01.009

**Table 1**
Data format for the $k$th occasion.

| | Response ($Y_k$) | | |
|---|---|---|---|
| Group ($X$) | $j = 1$ | $j = 2$ | |
| $i = 1$ | $n_{11.k}$ | $n_{12.k}$ | $n_1$ |
| $i = 2$ | $n_{21.k}$ | $n_{22.k}$ | $n_2$ |

However, correlated responses do not occur only in the repeated measurements context. It can be the case that each subject contributes two measurements; for example, for each eye or ear. In this framework, Tang et al. (2008) compared various test procedures for testing the equality of proportions of correlated binary responses.

All the above-mentioned references address their problem for cases with individual based information. We deal with the case where the individual level information is not available and thus all the above methods are not applicable. This situation is common in the framework of aggregated health data, in meta-analytic procedures and in official statistics. In such cases, the traditional and most common form of reporting data is through marginal tables (Fienberg and Slavkovic, 2004). We consider the fundamental case of a repeated binary response measured twice on samples of two independent groups, formed by a binary exposure. The only available data are the two marginal $2 \times 2$ tables, cross-classifying subjects by exposure and response category, one for each occasion. In this context, we would like to estimate the association between exposure (or prognostic factor) and outcome at each time point and to evaluate whether this association remains constant over time.

In the case of several independent $2 \times 2$ tables, the most well-known test of constant association across these tables, is that of Mantel–Haenszel. However, this approach could not be adopted in our framework, since it is not valid in the case of intraclass and/or interclass correlation. Most of the modifications of the Mantel–Haenszel test that appear in the literature deal with the intraclass correlation. An exception is a procedure given by Begg (1999), which accounts for dependence within and between strata. However, her variance correction factor is based on the GEE independent equations. Liao (1986) proposed a hierarchical Bayesian model for multiple $2 \times 2$ tables, under which the tables borrow information from each other. However, Liao's model is not fully Bayesian, since the nuisance parameters are eliminated by conditioning instead of integration.

Under this framework of binary correlated data, the correlation that exists between the two marginal tables needs special treatment and we have to account for it in our procedure. The unobserved or missing items' information and thus the missing transition probability matrix will be recovered through latent variables in the framework of a Bayesian analysis.

The paper is organized as follows. Section 2 describes our setup and the associated data formulation. Furthermore, it explains the way the summary information given at the two occasions is joined. Section 3 describes the Bayesian procedure for the parameter estimation and introduces the latent data that are applied in order to handle the missing information. Further on, the required Monte Carlo (MC) simulation is described while the hypothesis of constant exposure effect over time is treated through model diagnostics based on predictive $p$-values. In Section 4 we illustrate the proposed method via an example, common in the literature on longitudinal models. Finally, results are summarized in Section 5.

## 2. Model formulation

Consider a binary characteristic (response) measured successively at two time points ($k = 1, 2$) for two independent groups. For each time $k$ the data are presented in the form of a $2 \times 2$ contingency table (see Table 1), where $n_{ij.k}$ represent the cell counts for group $i$ ($i = 1, 2$) and category response $j$ ($j = 1, 2$). Without loss of generality, we assume that the response $j = 1$ is the 'success' category. Thus, since we have two independent groups we assume that cell frequencies $n_{11.k}$ and $n_{21.k}$ are independently binomial distributed, i.e. $n_{11.k} \sim \mathcal{B}inomial(n_1, \pi_{11.k})$ and $n_{21.k} \sim \mathcal{B}inomial(n_2, \pi_{21.k})$. Primary object is to estimate the success probabilities $\pi_{11.k}$ and $\pi_{21.k}$ as well as the odds ratio

$$\theta_k = \frac{\pi_{11.k}\pi_{22.k}}{\pi_{12.k}\pi_{21.k}}, \quad k = 1, 2, \tag{2.1}$$

that compare the two groups at each occasion $k$ in terms of their binary response.

The success probabilities for the first measurement, $\pi_{11.1}$ and $\pi_{21.1}$, and consequently the odds ratio $\theta_1$, can be estimated directly using the corresponding cell frequencies. However, the estimation of the success probabilities for the second table, $\pi_{11.2}$ and $\pi_{21.2}$, need to take into consideration not only the cell frequencies of this table but also the underlying correlation between probabilities $\pi_{i1.1}$ and $\pi_{i1.2}(i = 1, 2)$. This correlation affects also the estimation of the odds ratio at the second marginal table, i.e. $\theta_2$.

Let $w_{ij}$ denote the conditional probability of a subject of the $i$th group to remain in the same response category at the second occasion, i.e.

$$w_{ij} = P(Y_2 = j | X = i,\ Y_1 = j).$$

Then, the probabilities for the second table are given by $\mathbf{\Pi}_2 = \mathbf{W} \cdot \mathbf{\Pi}_1$, where

$$\mathbf{\Pi}_k = (\pi_{11.k}, \pi_{12.k}, \pi_{21.k}, \pi_{22.k})', \quad k = 1, 2,$$

**Table 2**
Transition matrix for group $i$.

| | $k=2$ | | |
|---|---|---|---|
| $k=1$ | $Y_2=1$ | $Y_2=2$ | |
| $Y_1=1$ | $z_i$ | $n_{i1.1}-z_i$ | $n_{i1.1}$ |
| $Y_1=2$ | $n_{i1.2}-z_i$ | $z_i+n_i-n_{i1.1}-n_{i1.2}$ | $n_i-n_{i1.1}$ |
| | $n_{i1.2}$ | $n_i-n_{i1.2}$ | $n_i$ |

and

$$\mathbf{W}=\begin{bmatrix} w_{11} & 1-w_{12} & 0 & 0 \\ 1-w_{11} & w_{12} & 0 & 0 \\ 0 & 0 & w_{21} & 1-w_{22} \\ 0 & 0 & 1-w_{21} & w_{22} \end{bmatrix}.$$

Using the conditional probabilities, the success probability at the second measurement for group $i$ is given by the equation

$$\pi_{i1.2}=w_{i1}\pi_{i1.1}+(1-w_{i2})(1-\pi_{i1.1}),\quad i=1,\ 2. \tag{2.2}$$

This equation implies that for each group, the success (or failure) probability at the second time point is a direct function of the success and failure probabilities of its previous classification state and the corresponding transition (conditional) probabilities $w_{i1}$ and $w_{i2}$. Thus, it can be also viewed as a weighted mean of the transition probabilities $w_{i1}$ and $w_{i2}$, with weights being the success and failure probabilities at the first time point. Under this setup, we allow the estimates of the success probabilities and the odds ratio of the second measurement to contain information from the first measurement, since the two tables are assumed correlated.

## 3. Bayesian inference

In order to deal with the missing information, that is the item's information, we use a data augmentation scheme. Given the number of individuals $n_i$ assigned at each group $i$ ($i=1,2$), the observed data are $(n_{11.k}, n_{21.k};\ k=1,2)$. In addition, and for each group $i$, the latent data $z_i$ are introduced. This discrete latent variable denotes the unknown number of subjects of group $i$ who remained in the success category at both time points. The transition frequency scheme between first and second measurement of the subjects assigned in group $i$ is provided in Table 2. In this Table, the marginal frequencies are the only available information while the missing cells must be estimated via the estimation of the latent variable $z_i$. To solve the problem and handle both, observed and unobserved variables, we developed a data augmentation algorithm (Tanner and Wong, 1987; Liu, 2001, section 6.4).

Under the above augmentation scheme, the full model likelihood can be written as

$$\begin{aligned} f(data,\mathbf{z}|\vartheta,n_1,n_2) &= \prod_{i=1}^{2} f(n_{i1.1},n_{i1.2},z_i|\vartheta_i,n_i) \\ &= \prod_{i=1}^{2} f(n_{i1.1}|\vartheta_i,n_i)f(z_i|n_{i1.1},\vartheta_i,n_i)f(n_{i1.2}|n_{i1.1},\vartheta_i,n_i), \end{aligned}$$

where $\vartheta_i=(\pi_{i1.1},w_{i1},w_{i2})$ is the parameter vector of the model for group $i$, $\vartheta=(\vartheta_1,\vartheta_2)$ is the full parameter vector and $data=(n_{i1.k};\ i,k=1,2)$ is the available observed data. Note that $n_{i2.k}$ is expressed as a function of the above observed data since $n_{i2.k}=n_i-n_{i1.k}$.

In the above decomposition, the distribution of $n_{i1.1}$, given the parameter vector $\vartheta_i$ and the group total $n_i$, is simply a Binomial distribution with success probability $\pi_{i1.1}$:

$$n_{i1.1}|\vartheta_i,n_i\sim\mathcal{B}inomial(n_i,\pi_{i1.1})\ ,$$

while the conditional distribution of the latent variable $z_i$ given the frequency $n_{i1.1}$, is also a binomial distribution

$$z_i|n_{i1.1},\vartheta_i,n_i\sim\mathcal{B}inomial(n_{i1.1},w_{i1}).$$

Analogously, the resulting conditional distribution of $n_{i1.2}$, given the latent data $z_i$ and $n_{i1.1}$, is a shifted Binomial distribution with $z_i\le n_{i1.2}\le n_i-n_{i1.1}+z_i$ since

$$n_{i1.2}-z_i|z_i,n_{i1.1},\vartheta_i,n_i\sim\mathcal{B}inomial(n_i-n_{i1.1},1-w_{i2})\ .$$

Because the two groups are assumed to be independent, the full likelihood of the augmented data is

$$\begin{aligned} f(data,\mathbf{z}|\vartheta) &= \prod_{i=1}^{2} f(n_{i1.1}|\vartheta_i,n_i)f(z_i|n_{i1.1},\vartheta_i,n_i)f(n_{i1.2}|n_{i1.1},\vartheta_i,n_i) \\ &= \prod_{i=1}^{2}\left\{f_{\mathcal{B}}(n_{i1.1};\ n_i,\pi_{i1.1})f_{\mathcal{B}}(z_i;\ n_{i1.1},w_{i1})f_{\mathcal{B}}(n_{i1.2}-z_i;\ n_i-n_{i1.1},1-w_{i2})\right\}, \end{aligned}$$

where $f_{\mathcal{B}}(x;\ n, \pi)$ is the probability function of the $\mathcal{B}inomial(n, \pi)$ distribution. Thus,

$$\begin{aligned} f(data, \mathbf{z}|\vartheta) &= \prod_{i=1}^{2} f(data, z_i|\vartheta) \\ &= \prod_{i=1}^{2} \{C(n_i, n_{i1.1}, n_{12.1}, z_i)\, \pi_{i1.1}^{n_{i1.1}}\, (1-\pi_{i1.1})^{n_i - n_{i1.1}}\, w_{i1}^{z_i}\, (1-w_{i1})^{n_{i1.1}-z_i}\, w_{i2}^{z_i+n_{i2.1}-n_{i1.2}} \\ &\quad \times\ (1-w_{i2})^{n_{i1.2}-z_i}\, I(z_i^{\min} \le z_i \le z_i^{\max})\}, \end{aligned} \tag{3.3}$$

where $I(x)$ is the indicator function taking value equal to one if $x$ is true and zero otherwise, $z_i^{\min} = \max\{0, n_{i1.1} + n_{i1.2} - n_i\}$, $z_i^{\max} = \min\{n_{i1.1}, n_{i1.2}\}$ and

$$\begin{aligned} C(n_i, n_{i1.1}, n_{i1.2}, z_i) &= \binom{n_i}{n_{i1.1}} \binom{n_{i1.1}}{z_i} \binom{n_i - n_{i1.1}}{n_{i1.2} - z_i} \\ &= \frac{n_i!}{z_i!(n_{i1.1} - z_i)!(n_{i1.2} - z_i)!(n_i - n_{i1.1} - n_{i1.2} + z_i)!}. \end{aligned}$$

The priors imposed on the parameters of interest $\vartheta_i = (\pi_{i1.1}, w_{i1}, w_{i2})$, $i = 1, 2$, are specified as independent beta distributions, which is the usual choice for binomial probabilities:

$$\pi_{i1.1} \sim \mathcal{B}eta(a_{i0}, b_{i0}), \qquad w_{i1} \sim \mathcal{B}eta(a_{i1}, b_{i1}), \qquad w_{i2} \sim \mathcal{B}eta(a_{i2}, b_{i2}). \tag{3.4}$$

Under this prior setup, the posterior distribution has the following structure

$$\begin{aligned} f(\vartheta, \mathbf{z}|data) &\propto \prod_{i=1}^{2} f(\vartheta_i, z_i|data) \\ &= \prod_{i=1}^{2} f(\pi_{i1.1}|data) f(w_{i1}|z_i, data) f(w_{i2}|z_i, data) f(z_i|data). \end{aligned} \tag{3.5}$$

For group $i$, the conditional posterior distribution $f(\vartheta_i, z_i|data)$ is defined through Eq. (3.5) as a product of the marginal posteriors of each component of $\vartheta_i$ and that of the latent $z_i$. It turns out that the marginal posterior distribution of the success probability $\pi_{i1.1}$ is the beta distribution

$$f(\pi_{i1.1}|data) = \mathcal{B}eta\,(n_{i1.1} + a_{i0},\ n_i - n_{i1.1} + b_{i0})\,. \tag{3.6}$$

The posterior distributions of the conditional probabilities $w_{i1}$ and $w_{i2}$, given the augmented data, are beta as well:

$$f(w_{i1}|z_i, data) = \mathcal{B}eta\,(z_i + a_{i1},\ n_{i1.1} - z_i + b_{i1})\,, \tag{3.7}$$

and

$$f(w_{i2}|z_i, data) = \mathcal{B}eta\,(z_i + n_{i2.1} - n_{i1.2} + a_{i2},\ n_{i1.2} - z_i + b_{i2})\,. \tag{3.8}$$

Finally, the distribution of the latent data $z_i$ given the observed data $(n_{i1.1}, n_{i1.2})$ and the group total $n_i$ appearing as the last term of (3.5) is given by integrating out the remaining parameters from the joint posterior as follows,

$$\begin{aligned} f(z_i|data) &= \int f(\pi_{i1.1}, w_{i1}, w_{i2}, z_i|data)\mathrm{d}\pi_{i1.1}\mathrm{d}w_{i1}\mathrm{d}w_{i2} \\ &= \int \frac{f(data, z_i|\pi_{i1.1}, w_{i1}, w_{i2}) f(\pi_{i1.1}) f(w_{i1}) f(w_{i2})}{f(data)} \mathrm{d}\pi_{i1.1}\mathrm{d}w_{i1}\mathrm{d}w_{i2} \\ &\propto \frac{\Gamma(z_i + a_{i1})}{\Gamma(z_i + 1)} \frac{\Gamma(n_{i1.1} - z_i + b_{i1})}{\Gamma(n_{i1.1} - z_i + 1)} \frac{\Gamma(z_i + n_i - n_{i1.1} - n_{i1.2} + a_{i2})}{\Gamma(z_i + n_i - n_{i1.1} - n_{i1.2} + 1)} \\ &\quad \times \frac{\Gamma(n_{i1.2} - z_i + b_{i2})}{\Gamma(n_{i1.2} - z_i + 1)} I(z_i^{\min} \le z_i \le z_i^{\max})\,. \end{aligned} \tag{3.9}$$

The marginal distribution of the latent data depends only on the observed cell frequencies and not on model parameters. This allows us to estimate the mean value of the number of subjects that remains in success using standard simulation methods for discrete variables. Also, even though the marginal posterior distributions $f(w_{ij}|data)$ $(i, j = 1, 2)$ of the conditional probabilities are not of a standard form, it is easy to calculate analytically their moments in closed-form expressions using the posterior means of the latent data. More specific, using the posterior distributions of the conditional probabilities given in (3.7) and (3.8), the first moment of the marginal posterior distribution of each $w_{ij}$ $(i, j = 1, 2)$ is

$$E(w_{i1}|data) = E\,\{E\,(w_{i1}|z_i, data)\} = \frac{E(z_i) + a_{i1}}{n_{i1.1} + a_{i1} + b_{i1}},$$

and

$$E(w_{i2}|data) = E\{E(w_{i2}|z_i, data)\} = \frac{E(z_i) + n_{i2.1} - n_{i1.2} + a_{i2}}{n_{i2.1} + a_{i2} + b_{i2}},$$

where $E(z_i)$ is the posterior mean of the latent variable $z_i$. The above expressions give estimations of the mean values for the conditional probabilities $w_{ij}$ without using an iterative scheme. Using the same setup we can estimate the second central moment, i.e. the variance, of the marginal posterior distribution of $w_{i1}$ or $w_{i2}$ $(i = 1, 2)$:

$$\begin{aligned} Var(w_{i1}|data) &= E\{Var(w_{i1}|z_i, data)\} + Var\{E(w_{i1}|z_i, data)\} \\ &= \frac{E(z_i)(n_{i1.1} + b_{i1} - a_{i1}) - E(z_i^2) + a_{i1}b_{i1} + a_{i1}n_{i1.1}}{(n_{i1.1} + a_{i1} + b_{i1})^2(n_{i1.1} + a_{i1} + b_{i1} + 1)}, \end{aligned}$$

and

$$\begin{aligned} Var(w_{i2}|data) &= E\{Var(w_{i2}|z_i, data)\} + Var\{E(w_{i2}|z_i, data)\} \\ &= \frac{E(z_i)(b_{i2} + n_{i1.2} - a_{i2}) - E(z_i^2) + a_{i2}b_{i2}}{(n_{i2.1} + a_{i2} + b_{i2})^2(n_{i2.1} + a_{i2} + b_{i2} + 1)} + \frac{n_{i2.1}(b_{i2} + a_{i2}) - n_{i1.2}(n_{i2.1} + b_{i2})}{(n_{i2.1} + a_{i2} + b_{i2})^2(n_{i2.1} + a_{i2} + b_{i2} + 1)}. \end{aligned}$$

Because we are interested in the estimation of the success probabilities and odds ratios for the two time points, we must use an iterative scheme to simulate both latent and parameter values. Producing simulated values of the marginal distribution of the $z_i$'s (3.9) we can use Monte Carlo method to generate samples of the posterior distributions of model parameters $\vartheta$. This method is described in the following subsection.

### 3.1. Monte Carlo simulation

Monte Carlo simulation (Geweke, 1989) is used to generate the parameters of interest $\vartheta$ from the corresponding conditional posterior distributions $f(\pi_{i1.1}|data)$ and $f(w_{ij}|z_i, data)$ $(i, j = 1, 2)$ through the generation of an i.i.d sample of the latent data $\mathbf{z}$ from (3.9). Hence we adopt the following sampling scheme for $t = 1, 2, \ldots, T$ iterations

(1) Sample $z_i^{(t)}$, $i = 1, 2$ from (3.9).
(2) Sample $\pi_{i1.1}^{(t)}$, $w_{i1}^{(t)}$ and $w_{i2}^{(t)}$ for $i = 1, 2$ from the posterior distributions (3.6), (3.7) and (3.8) respectively.
(3) Calculate $\pi_{i1.2}^{(t)} = w_{i1}^{(t)}\pi_{i1.1}^{(t)} + \left(1 - w_{i2}^{(t)}\right)\left(1 - \pi_{i1.1}^{(t)}\right)$.
(4) Calculate the odds ratios $\theta_k^{(t)}$ for $k = 1, 2$ by (2.1).

The within brackets superscript in the above sampling scheme is used to denote the value of the underlined parameter generated in the $t$ iteration of the algorithm.

### 3.2. Model diagnostics

In order to evaluate the plausibility of certain assumptions concerning the estimated odds ratios we compute posterior predictive $p$-values (Meng, 1994).

In our context, the hypotheses of special interest are

$$H_1 : \theta_1 = 1, \qquad H_2 : \theta_2 = 1 \quad \text{and} \quad H_3 : \theta_1 = \theta_2. \tag{3.10}$$

To test the above hypothesis we compute the predictive distribution $f(\mathbf{n}^{pred}|data)$; where $\mathbf{n}^{pred} = (n_{i1.k}^{pred},\ i, k = 1, 2)$, are the predictive data. In order to test the above hypotheses we calculate the odds ratios $\widehat{\theta}_k^{pred}$ $(k = 1, 2)$

$$\widehat{\theta}_k^{pred} = \frac{n_{11.k}^{pred}\left(n_2 - n_{21.k}^{pred}\right)}{n_{21.k}^{pred}\left(n_1 - n_{11.k}^{pred}\right)}, \tag{3.11}$$

based on $\mathbf{n}^{pred}$ and their corresponding ratio $\widehat{\theta}_2^{pred}/\widehat{\theta}_1^{pred}$.

Hence we wish to estimate the posterior predictive distributions

$$f\left(\widehat{\theta}_k^{pred}|data\right),\ k = 1, 2 \quad \text{and} \quad f\left(\widehat{\theta}_2^{pred}/\widehat{\theta}_1^{pred}|data\right),$$

and the following $p$-values

$$p_k = P\left(\widehat{\theta}_k^{pred} > 1|data\right) \quad \text{for } k = 1, 2 \quad \text{and} \quad p_3 = P\left(\widehat{\theta}_2^{pred}/\widehat{\theta}_1^{pred} > 1|data\right), \tag{3.12}$$

for testing the corresponding hypotheses $H_\xi$ for $\xi = 1, 2, 3$. Values close to 0.5 indicate that the assumed hypothesis are plausible while values to the extremes indicate that these hypotheses are not plausible under the assumed model and the observed data.

Estimation of the above $p$-values can be achieved by using simple additional steps in the above Monte Carlo algorithm. To be more specific, we only need to generate replicated/predictive values $\mathbf{n}^{pred\,(t)}$ from the corresponding model likelihood and then calculate the estimated odds $\widehat{\theta}_k^{pred\,(t)}$ (Meng, 1994). Hence in the steps of the algorithm introduced in the previous subsection, we must add

5. For $i, k = 1, 2$, generate $n_{i1.k}^{pred} \sim f\left(\mathbf{n}^{pred}|\mathbf{z}, \vartheta^{(t)}\right) \propto f\left(\mathbf{n}^{pred}, \mathbf{z}|\vartheta^{(t)}\right)$, as described in what follows.
6. For $k = 1, 2$, calculate $\widehat{\theta}_k^{pred\,(t)}$ from (3.11) using $\mathbf{n}^{pred\,(t)}$.
7. Set binary indicators $\omega_k^{(t)} = I\left(\widehat{\theta}_k^{pred\,(t)} > 1\right)$ for $k = 1, 2$ and $\omega_3^{(t)} = I\left(\widehat{\theta}_2^{pred\,(t)} > \widehat{\theta}_1^{pred\,(t)}\right)$.

To generate the replicated data in step 5 described above, we need to consider the likelihood function (3.3). The marginal distribution of $n_{i1.1}$ and $n_{i1.2}$ (for group $i = 1, 2$) is given by

$$f(n_{i1.1}, n_{i1.2}|z_i, \vartheta_i) = f(n_{i1.1}|z_i, \vartheta_i) \cdot f(n_{i1.2}|n_{i1.1}, z_i, \vartheta_i),$$

where

$$\begin{aligned} f(n_{i1.1}|z_i, \vartheta_i) &\propto \frac{\left[\pi_{i1.1}(1-w_{i1})\right]^{n_{i1.1}}(1-\pi_{i1.1})^{n_i-n_{i1.1}}}{(n_{i1.1}-z_i)!(n_i-n_{i1.1})!} I(z_i \le n_{i1.1}) \\ &= f_{\mathcal{B}}\left(n_{i1.1}-z_i; n_i-z_i, \frac{\pi_{i1.1}(1-w_{i1})}{\pi_{i1.1}(1-w_{i1})+(1-\pi_{i1.1})}\right), \end{aligned}$$

and

$$\begin{aligned} f(n_{i1.2}|n_{i1.1}, z_i, \vartheta_i) &\propto \frac{(n_i-n_{i1.1})!w_{i2}^{n_i-n_{i1.1}-n_{i1.2}}(1-w_{12})^{n_{i1.2}}}{(z_i+n_i-n_{i1.1}-n_{i1.2})!(n_{i1.2}-z_i)!} I(z_i \le n_{i1.2} \le z_i+n_i-n_{i1.1}) \\ &= f_{\mathcal{B}}(n_{i1.2}-z_i;\ n_i-n_{i1.1},\ 1-w_{i2}). \end{aligned}$$

Since $f(n_{i1.2}|n_{i1.1}, z_i, \vartheta_i)$ are the probability functions of binomial distributions shifted by $z_i$, we can generate $n_{i1.1}^{pred}$ and $n_{i1.2}^{pred}$ by the following steps

5a. For $i = 1, 2$, generate $v_{i1}^{(t)} \sim \mathcal{B}inomial\left(n_i - z_i^{(t)}, \frac{\pi_{i1.1}^{(t)}\left(1-w_{i1}^{(t)}\right)}{\pi_{i1.1}^{(t)}\left(1-w_{i1}^{(t)}\right)+\left(1-\pi_{i1.1}^{(t)}\right)}\right)$ and set $n_{i1.1}^{pred\,(t)} = v_{i1}^{(t)} + z_i^{(t)}$.

5b. For $i = 1, 2$, generate $v_{i2}^{(t)} \sim \mathcal{B}inomial\,(n_i - n_{i1.1},\ 1-w_{i2})$ and set $n_{i1.2}^{pred\,(t)} = v_{i2}^{(t)} + z_i^{(t)}$.

After generating $\omega_k^{(t)}$ as described above, the corresponding $p$-values are estimated by

$$\widehat{p}_\xi = \frac{1}{T}\sum_{t=1}^{T}\omega_\xi^{(t)} \quad \text{for } \xi = 1, 2, 3,$$

which are simply the proportions of Monte Carlo generated values satisfying the conditions used in (3.12).

All the computations described above are implemented using R.

## 4. Illustrative example

To illustrate our procedure, we analyze a subset of a longitudinal study on the health effects of air pollution carried out at six cities (Ware et al., 1984). One of the objectives of this study was to determine how maternal smoking and outside air quality affects respiratory illness in children. This example is very popular in the literature and has been studied under various models for longitudinal data, as in Fitzmaurice and Laird (1993), who applied a multivariate logit model. For illustrative purposes, we have chosen an example for which the complete item information is available, so that we can evaluate our latent data approach compared to that based on full data. The data set contains complete records on 537 children from Ohio, each of whom was examined at ages 7 through 10. The wheeze status of a child (1 = yes, 0 = no) is the repeated binary response and maternal smoking (1 = mother was a regular smoker, 0 = otherwise) is treated as a fixed variable and forms the two independent groups of children. We will use the summary data for the first and last measurement at ages 7 and 10. Table 3 gives the two $2 \times 2$ tables of maternal smoking by wheeze status.

Because we assume prior ignorance on model parameters, we use the Jeffrey's non-informative priors given by

$$\pi_{i1.1} \sim \mathcal{B}eta\,(0.5, 0.5)\,, \qquad w_{i1} \sim \mathcal{B}eta\,(0.5, 0.5)\,, \quad \text{and} \quad w_{i2} \sim \mathcal{B}eta\,(0.5, 0.5)\,.$$

All results presented in the paper are based on 5000 iterations. Such a sample needed around 5.5 seconds to be generated by our R function on a desktop computer with a Pentium M 1.73 GHz processor and 504MB of RAM.

The posterior means of the parameter vector

$$\vartheta = (\vartheta_1, \vartheta_2) = (\pi_{11.1}, w_{11}, w_{12}, \pi_{21.1}, w_{21}, w_{22}),$$

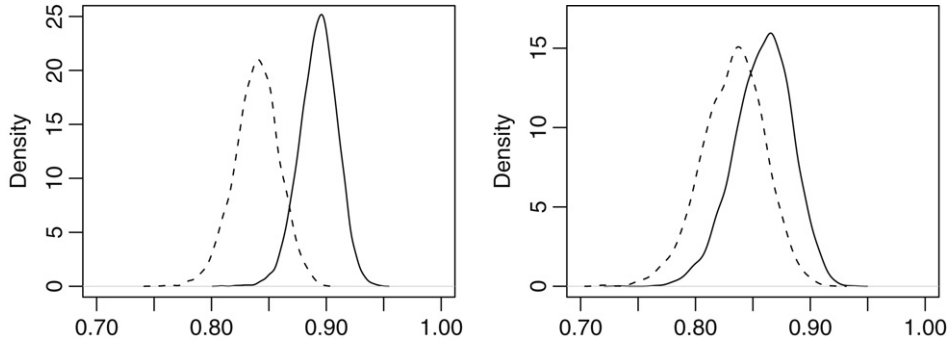
**Table 3**
Six cities data set for children at ages 7 and 10.

| Age | Maternal smoking | Wheeze status | | |
|---|---|---|---|---|
| | | No | Yes | |
| 7 years | No | 294 | 56 | 350 |
| ($k = 1$) | Yes | 156 | 31 | 187 |
| 10 years | No | 313 | 37 | 350 |
| ($k = 2$) | Yes | 161 | 26 | 187 |

**Table 4**
Posterior summaries of model parameters for the six cities data set.

| Parameter | Mean | SD[a] | 2.5% | Median | 97.5% | |
|---|---|---|---|---|---|---|
| $\pi_{11.1}$ | 0.839 | 0.019 | 0.799 | 0.840 | 0.876 | |
| $\pi_{21.1}$ | 0.832 | 0.027 | 0.777 | 0.834 | 0.882 | |
| $\pi_{11.2}$ | 0.892 | 0.016 | 0.857 | 0.893 | 0.923 | |
| $\pi_{21.2}$ | 0.858 | 0.025 | 0.806 | 0.859 | 0.904 | |
| $w_{11}$ | 0.943 | 0.047 | 0.854 | 0.951 | 0.999 | |
| $w_{12}$ | 0.371 | 0.242 | 0.001 | 0.405 | 0.725 | |
| $w_{21}$ | 0.932 | 0.062 | 0.804 | 0.949 | 0.999 | |
| $w_{22}$ | 0.508 | 0.296 | 0.003 | 0.578 | 0.898 | |
| Latent data | | | | | | |
| $z_1$ | 277.588 | 13.437 | 257 | 280 | 294 | |
| $z_2$ | 145.746 | 9.227 | 130 | 148 | 156 | |
| Estimated ratio | | | | | | $P(ratio > 1\|data)$ |
| $\pi_{11.2}/\pi_{11.1}$ | 1.064 | 0.025 | 1.015 | 1.063 | 1.117 | 0.993 |
| $\pi_{21.2}/\pi_{21.1}$ | 1.032 | 0.034 | 0.962 | 1.030 | 1.106 | 0.863 |

[a] SD = Standard deviance.




(a) Non-smoker mother. (b) Smoker mother.

**Fig. 1.** Posterior densities of the estimated no respiratory illness probabilities at age 7 ($\cdots$) and at age 10 ($-$).

and the estimated success probabilities $\pi_{i1.2}$ of no respiratory illness at age 10 for group $i = 1, 2$ are given in Table 4. In this table we also give summaries for the estimated posterior ratios $\pi_{i1.2}/\pi_{i1.1}$ of the disease-free probabilities for each group, comparing the two time points. We also give summary results of the latent data $z_i$ ($i = 1, 2$), that is the number of subjects of group $i$ that remained in success at the second measurement. We see that the probability of a child to be disease-free at the age of 7 is about 84% for both groups defined by the maternal smoking habit. At age 10 the probability of no wheeze increases to 89% and 86% for a non-smoker and a regular smoker mother respectively. Maternal smoking is thus a negative prognostic factor on respiratory illness of children over time. This is confirmed by the large values 0.993 and 0.863 of the estimated posterior probabilities $P\left(\frac{\pi_{i1.2}}{\pi_{i1.1}} > 1|data\right)$ for $i = 1, 2$ respectively. Finally, only 16 children of the non-smoker maternity group out of 294 healthy children at age 7, developed respiratory illness at the age of 10. For the smoker maternity group 10 children of negative diagnosis at the age of 7 out of 156 children, developed the disease at age 10. The true values of the latent data extracted from the full data set given in Fitzmaurice and Laird (1993) are 275 for $z_1$ and 143 for $z_2$, very close to the estimated mean of our latent data.

The posterior densities of the estimated disease-free probabilities at both ages for each group of children are depicted in Fig. 1. From the posterior density plots, it is clear to observe a higher increase of the disease-free probability over time for the non-smoking maternity group compared to the corresponding lower increase for the smoking maternity group.

Posterior summaries of the odds ratios for the two time points and the estimated ratio $\theta_2/\theta_1$ are provided in Table 5. In Fig. 2 we also give the posterior densities of the odds ratios comparing ages 7 and 10. At age 7 the odds of no-illness

**Table 5**
Posterior summaries of odds ratios $\theta_1$ and $\theta_2$.

| Parameter | Mean | SD[a] | 2.5% | Median | 97.5% | $P$(parameter $>1$\|*data*) |
|---|---|---|---|---|---|---|
| $\theta_1$ | 1.08 | 0.27 | 0.65 | 1.05 | 1.67 | 0.573 |
| $\theta_2$ | 1.41 | 0.39 | 0.79 | 1.36 | 2.34 | 0.875 |
| $\theta_2/\theta_1$ | 1.36 | 0.39 | 0.74 | 1.30 | 2.34 | 0.841 |

[a] SD = Standard deviance.

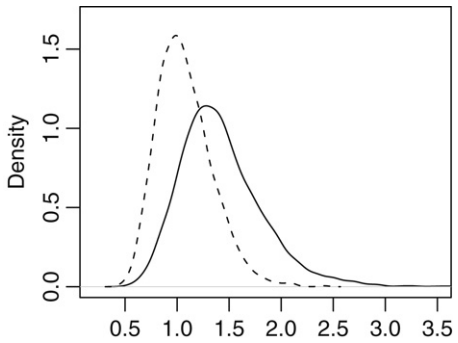



**Fig. 2.** Posterior densities of odds ratios at age 7 ($\cdots$) and age 10 ($-$).

**Table 6**
Estimated descriptive measures of predictive counts (replicated data).

| Parameters | Mean | SD[a] | 2.5% | Median | 97.5% | Observed data |
|---|---|---|---|---|---|---|
| $n_{11.1}^{pred}$ | 294.2 | 4.8 | 284 | 294 | 305 | 294 |
| $n_{12.1}^{pred}$ | 55.8 | 4.8 | 45 | 56 | 66 | 56 |
| $n_{21.1}^{pred}$ | 156.2 | 3.6 | 149 | 156 | 164 | 156 |
| $n_{22.1}^{pred}$ | 30.8 | 3.6 | 23 | 31 | 38 | 31 |
| $n_{11.2}^{pred}$ | 312.7 | 5.9 | 301 | 313 | 325 | 313 |
| $n_{12.2}^{pred}$ | 37.3 | 5.9 | 25 | 37 | 49 | 37 |
| $n_{21.2}^{pred}$ | 160.9 | 4.2 | 153 | 161 | 170 | 161 |
| $n_{22.2}^{pred}$ | 26.0 | 4.2 | 17 | 26 | 34 | 26 |
| $\theta_1^{pred}$ | 1.053 | 0.190 | 0.712 | 1.043 | 1.487 | |
| $\theta_2^{pred}$ | 1.404 | 0.386 | 0.787 | 1.364 | 2.296 | |

[a] SD = Standard deviance.

is close for the two groups since the posterior probability $P(\theta_1 > 1|data)$ is 0.573. In contrast, at age 10, the estimated odds of no illness for children with a non-smoker mother is 1.4 times the corresponding odds for children with smoking mother since the posterior probability $P(\theta_2 > 1|data)$ is 0.875. This also indicates that mother's smoking habit influences the development of respiratory illness in children over time.

An estimate of the ratio $\theta_2/\theta_1$ can be extracted by the analysis of Fitzmaurice and Laird (1993) based on the complete individual level data. Following their results, $\widehat{\theta_2}/\widehat{\theta_1} = e^{0.0711(10-7)} = 1.24$ indicating an increase of the odds ratio at age 10 compared to age 7 by 24%. This estimate is close to the posterior mean of the ratio $\theta_2/\theta_1 = 1.36$ (see Table 5) when only the marginal information of each table is available.

We performed a sensitivity analysis by setting $a_{ik} = b_{ik} = c$ ($k = 0, 1, 2, i = 1, 2$) for the corresponding *Beta* priors, defined in (3.4), and for values $c = 1/10,\ 1/5,\ 1/3,\ 1/2,\ 1$. We observed minor differences in the posterior summaries for the $\pi_{ij.k}$'s and the $\theta_k$'s as well as their ratio. Note that the differences are slight bigger for the $w_{ij}$'s, fact that should be expected, since these are latent variables and thus are more sensitive in prior information. However, the differences on the $\theta_k$'s, which are the parameters of interest, are minor.

To complete the analysis, we proceed with the predictive analysis providing summaries for the predictive/replicated data generated as described in Section 3.2 (see Table 6). As expected, estimated means of the cell frequencies at both ages are close to the observed frequencies. Moreover, the calculated odds ratios based on the replicated data are of the same magnitude as the posterior odds given in Table 5 with slightly lower the odds ratio of the first time measurement.

Note that the assumed model imposes a saturated structure on each of the two $2 \times 2$ marginal tables we observe. Thus, it is reasonable that the medians of their predictive counts coincide with the observed counts (data), as in our example, see Table 6.

We further examine the plausibility of hypotheses $H_1$, $H_2$ and $H_3$ given in (3.10) using the posterior predictive $p$-values (3.12). For the first hypothesis the predictive $p$-value was estimated equal to 0.622. Hence the assumed null hypothesis is plausible since the value of $\theta_1 = 1$ assuming independence between mother's smoking status and child's wheezing problems lies in the close to the center of the posterior distribution $f(\theta_1|data)$. For the next two hypotheses, the corresponding predictive $p$-values were estimated equal to 0.887 and 0.776 respectively. Hence, the second posterior $p$-value (0.89) indicates a moderate association between mother's smoking status and child's wheezing problems at age 10. Finally, for $H_3$ the corresponding estimated $p$-value indicates a mild increase of the mother's smoking influence on the health of their child that need to be further investigated by considering additional time points.

## 5. Discussion and conclusion

In this paper we developed a Bayesian approach for analyzing two correlated binary responses, measured for two independent groups. The two $2 \times 2$ cross-tabulations are marginal tables of the complete individual information at the two time points. These tables contain no information on the transition frequencies, the way that subjects moved between the two response categories of these two measurements. This context is common in aggregated data and in meta-analytic applications where recapture of the complete data set allows more complex analytic methods. In order to handle the missing item information we introduced latent variables in a data augmentation scheme. We modeled the two tables using a simple Monte Carlo simulation method and we estimated the missing data as well as the conditional probabilities of subjects to remain in the same response category at the second measurement. Then we estimated the cell probabilities and the corresponding odds ratios comparing the two groups in terms of the binary response. The correlation between the two time tables was induced in the estimation of the second's table cell probabilities via the estimation of the conditional probabilities for subjects of group $i$ ($i = 1, 2$) to remain in the same response category $j$ ($j = 1, 2$) at the second measurement (i.e. the $w_{ij}$'s). In order to test for associations between exposure and outcome at each measurement and also to test the hypothesis of constant exposure effect over time we produced the analogous predictive $p$-values. The proposed method was illustrated via a popular example in the literature of longitudinal studies. It is the subject of our current research to extend these results for the case of more than two occasions.

## Acknowledgements

The authors wish to thank the Editor and the anonymous referees for their productive comments which considerably improved the paper.

## References

Agresti, A., Klingenberg, B., 2005. Multivariate tests comparing binomial probabilities, with application to safety studies for drugs. Applied Statistics 54, 691–706.
Begg, M.D., 1999. Analyzing $k$ ($2 \times 2$) tables under cluster sampling. Biometrics 55, 302–307.
Chib, S., Carlin, B., 1999. On MCMC sampling in hierarchical longitudinal models. Statistics and Computing 9, 17–26.
Fienberg, S.E., Slavkovic, A.B., 2004. Making the release of confidential data from multi-way tables count. Chance 17, 5–10.
Fitzmaurice, G.M., Laird, N.M., 1993. A likelihood-based method for analyzing longitudinal binary responses. Biometrika 80, 141–151.
Geweke, J., 1989. Bayesian inference in econometric models using Monte Carlo integration. Econometrica 57, 1317–1339.
Jara, A., Garcia-Zattera, M.J., Lesaffre, E., 2007. A Dirichlet process mixture model for the analysis of correlated binary responses. Computational Statistics & Data Analysis 51, 5402–5415.
Liang, K.Y., Zeger, S.L., 1986. Longitudinal data analysis using generalized linear models. Biometrika 73, 13–22.
Liao, J.G., 1986. A hierarchical Bayesian model for combining multiple $2 \times 2$ tables using conditional likelihoods. Biometrics 55, 268–272.
Liu, J.S., 2001. Monte Carlo Strategies in Scientific Computing. Springer, New York.
Meng, X.L., 1994. Posterior predictive $p$-values. The Annals of Statistics 22, 1142–1160.
Tan, M., Tian, G.L., Ng, K.W., 2006. Hierarchical models for repeated binary data using the IBF sampler. Computational Statistics & Data Analysis 50, 1272–1286.
Tang, N.S., Tang, M.L., Qiu, S.F., 2008. Testing the equality of proportions for correlated otolaryngologic data. Computational Statistics & Data Analysis 52, 3719–3729.
Tanner, M.A., Wong, W.H., 1987. The calculation of posterior distributions by data augmentation. Journal of the American Statistical Association 82, 528–540.
Ware, J.H., Dockery, D.W., Spiro III, A., Speizer, F.E., Ferris, B.G., 1984. Passive smoking, gas cooking and respiratory health in children living in six cities. The American Review of Respiratory Disease 129, 366–374.