Journal of the
Royal Statistical Society

SERIES C
Applied
Statistics

Appl. Statist. (2009)
58, *Part* 2, *pp.* 237–246

Using Bayesian *p*-values in a 2 × 2 table of matched pairs with incompletely classified data

Yan Lin,

University of Texas M. D. Anderson Cancer Center, Houston, USA

Stuart Lipsitz,

Brigham and Women's Hospital, Boston, and Medical University of South Carolina, Charleston, USA

Debajyoti Sinha,

Florida State University, Tallahassee, USA

Atul A. Gawande,

Brigham and Women's Hospital, Boston, USA

Scott E. Regenbogen

Massachusetts General Hospital, Boston, USA

and Caprice C. Greenberg

Brigham and Women's Hospital, Boston, USA

[Received July 2006. Final revision August 2008]

Summary. Altham proposed Bayesian *p*-values for the analysis of a 2 × 2 contingency table that is formed from matched pairs. Using the same Bayesian perspective, we develop an extension of Altham's Bayesian *p*-values to a 2 × 2 table from matched pairs with missing data that are missing at random. The approach is applied to a rater agreement study, in which two surgeon–reviewers rated whether or not there was a communication breakdown in malpractice cases. We also use a simulation study to explore the power and type I error rate of the Bayesian *p*-values.

Keywords: Dirichlet prior; Exact McNemar test; Ignorable missing data

1. Introduction

Comparison of two correlated proportions or percentages is common in the biological and medical sciences. Correlated proportions typically arise from longitudinal studies, rater reliability studies, pre–post test designs and matched case–control studies, and such data can be displayed in a 2 × 2 contingency table. McNemar (1947) derived a χ^2-test statistic for the null hypothesis that the expected value of the difference between the two correlated proportions is 0. When the cell sizes in the 2 × 2 contingency table are small, an exact *p*-value for McNemar's test statistic

Address for correspondence: Yan Lin, Department of Biostatistics and Applied Mathematics, University of Texas M. D. Anderson Cancer Center, Houston, TX 77030, USA.
E-mail: hyanlin@mdanderson.org

0035–9254/09/58237

was developed by Mosteller (1952) and is based on the binomial distribution that is formed by conditioning on the discordant pairs of the 2 × 2 table. Unfortunately, besides small cell sizes, a frequent additional complication which is found in many studies is missing data; for example, either the row or column variable of the 2 × 2 table is missing data for some of the experimental units.

With missing data that are missing completely at random (Rubin, 1976), the exact McNemar *p*-value that is calculated after throwing out the missing data (i.e. by using only complete cases) will have type I error consistent with the nominal level. However, if the data are missing at random (Rubin, 1976), the exact McNemar *p*-value that is calculated after throwing out the missing data could have type I error higher than the nominal level. An extension of the exact McNemar *p*-value to the missing data instance with small samples would be appropriate but has not yet been developed.

In this paper, we propose an extension of the correlated proportions Bayesian *p*-value approach that was proposed by Altham (1971) to 2 × 2 tables with missing data. The Bayesian *p*-value is a tail probability based on a posterior distribution and was discussed by Altham (1969, 1971) and Casella and Berger (1987). Our extension of Altham's Bayesian approach uses different Dirichlet prior distributions for the cell probabilities of the 2 × 2 table, and the likelihood is based on all data, not just the subset of data in which the outcome for both pairs is observed. We consider four different specifications of the parameters of the Dirichlet prior distribution; with no missing data, one set of parameters of the Dirichlet prior leads to a Bayesian *p*-value which is identical to the exact McNemar *p*-value. The use of different priors can be considered as a sensitivity analysis to see whether the type I error and power change are sensitive to the different priors.

We also discuss both one- and two-sided *p*-values. For one-sided hypotheses, the Bayesian *p*-value is the posterior probability of the null hypothesis and equals the tail probability of a posterior distribution. For two-sided hypotheses, the posterior probability of the null hypothesis is a point probability, and thus not appropriate to use (see, for example, Berger and Sellke (1987) and Berger and Delampady (1987)). Instead, we calculate our two-sided *p*-value as two times the minimum of the two one-sided Bayesian *p*-values, as suggested in Fleiss *et al.* (2003).

Test statistics that have been proposed for matched pairs studies with missing data include the likelihood ratio statistic (Gokhale and Sirtonik, 1984) and the score statistic. For these test statistics, maximum likelihood estimates (MLEs) of the probabilities in the 2 × 2 table under the alternative and/or null hypothesis are needed but, with small samples (particularly 0 cell counts) and missing data such as ours, unique MLEs do not even exist for many data configurations (Fuchs, 1982), so neither the likelihood ratio statistic nor the score statistic can always be obtained. For 2 × 2 tables with missing data and a Dirichlet prior, the Bayesian posterior distribution always exists (although it may not be unimodal), even in small samples with 0 cell counts. We note that, when the Bayesian posterior distribution is multimodal, there may not be a non-unique MLE. However, since the Bayesian *p*-value is just a tail probability from a posterior distribution that always exists, the Bayesian *p*-value always exists, regardless of whether the Bayesian posterior distribution is unimodal.

In Section 2, we discuss our example. In Section 3, we specify probability models and priors. In Section 4, we analyse the data that are given in our example. In Section 5, we perform simulations comparing the *p*-values by using all data and just complete cases for the different Dirichlet priors.

2. Application: communication breakdowns in surgery

We analyse data from a study of medical malpractice claims in Section 4. The goal was to test whether two surgeon–reviewers had the same probability of rating 'hand-offs in care' as a

Table 1. Two surgeon reviews of malpractice claims data†

Reviewer 1's answer	*Results for reviewer 2's answers*			
	Yes	*No*	*Missing*	*Total*
Yes	26	1	2	29
No	5	18	9	32
Missing	4	4	0	8
Total	35	23	11	69

†Reviewers' answer to the question: was there a communication breakdown in the hand-off between physicians caring for the patient?

contributing factor in communication breakdowns in the malpractice claims (Greenberg *et al.*, 2007). Failures of communication are among the most common causes of serious adverse events in surgery (Rogers *et al.*, 2006). One of the potential factors in communication breakdowns is a hand-off in care between physicians caring for the patient. A hand-off is defined as the complete transfer of care from one physician to another when the first physician physically leaves the scene. In a previous study, Rogers *et al.* (2006) identified 258 malpractice claims from four liability insurers in which an identifiable error in care resulted in injury to a patient. To identify factors contributing to communication breakdowns and design interventions to prevent them, Greenberg *et al.* (2007) performed secondary review and analysis of 69 errors that were attributed to communication breakdowns.

Two surgeon–reviewers used a structured instrument to evaluate the 69 errors, and to identify important human and system factors contributing to the errors. Among many possible factors in communication breakdowns, each surgeon–reviewer was asked to determine whether a hand-off in care was associated with the communication breakdown (yes or no). As we see in Table 1, both surgeon–reviewers felt that a hand-off was involved in 26 out of the 69 breakdowns (37%). Unfortunately, eight reviews are missing for surgeon 1 and 11 reviews are missing for surgeon 2. In recent discussions with the two surgeon–reviewers who reviewed these malpractice claims 2 years earlier, they recalled that, with 27 questions on the instrument to rate, it was probably random oversight why the answers were missing. If this was so, then the missing data are missing completely at random.

The cell sizes in Table 1 (particularly the off-diagonal cells that were used in McNemar's test statistic) are small, meaning that we would like to use the exact p-value for McNemar's test statistic when testing whether the two raters have equal probability of finding problems with a hand-off of care.

Before analysing these data, it is of interest to explore the missing data mechanism that might be generating the missing data. A somewhat informal way (Chen and Little, 1999) to assess whether the data are missing completely at random (so that complete cases will give appropriate results) is to compare the marginal proportions in the 2×2 table of complete cases with the marginal proportions in those missing a row or column variable. In the complete cases, reviewer 1 rated 'yes' $27/50 = 54\%$ of the time and, in cases that were rated only by reviewer 1, reviewer 1 rated 'yes' only $2/11 = 18.2\%$. Fisher's exact test comparing these two proportions has a p-value of 0.046, which suggests that the data that were not reviewed by reviewer 2 may not be missing completely at random. In the complete cases, reviewer 2 rated 'yes' $31/50 = 62\%$ of the time and, in cases rated only by reviewer 2, reviewer 2 rated ‘yes’ $4/8=50\%$. Fisher’s exact test comparing these two proportions has a p-value of 0.70, which suggests that data that were not reviewed by reviewer 1 might be missing completely at random. Thus, because the data that were not reviewed by reviewer 2 may not be missing completely at random, it is possible that we may arrive at different conclusions by using all the data, and just the complete cases.

3. Models and Bayesian *p*-values

In a 2×2 table, let θ_{jk} denote the probability that the row variable equals j and the column variable equals k ($j=0,1$; $k=0,1$). For matched pairs data, the ‘row’ is the first subject in a pair and the ‘column’ is the second subject in the pair. The null hypothesis of interest is that the marginal probability that the outcome equals 1 is the same for both subjects in a pair, i.e. $H_0:\theta_{1+}=\theta_{+1}$, or, equivalently, $H_0:\theta_{10}=\theta_{01}$, where the ‘+’ subscript represents summation over the corresponding index for rows or columns.

When missing data occur, the observed data can be summarized in the form of a 2×2 table with supplemental margins, as shown in Table 2. In Table 2, Y_{jk} denotes the number of pairs who have both the row and the column variables observed, with response level j on the row variable and level k on the column variable ($j=0,1$; $k=0,1$). Also, Z_{j+} denotes the number of pairs with response level j on the row variable who are missing the column variable, and U_{+k} denotes the number of pairs with response level k on the column variable who are missing the row variable. We denote the observed data by $D=(Y_{11},Y_{10},Y_{01},Y_{00},Z_{1+},Z_{0+},U_{+1},U_{+0})$.

With no missing data, i.e. when $Z_{j+}=U_{+k}=0$ for all (j, k), and the observed data denoted by D_{c} consist only of $D_{\text{c}}=(Y_{11},Y_{10},Y_{01},Y_{00})$, the Y_{jk}s follow a multinomial distribution with cell probabilities θ_{jk}. With ignorable missing data (Rubin, 1976), the likelihood function that is based on all of the observed data D in Table 2 is proportional to

$$L(\theta|D)\propto\theta_{11}^{y_{11}}\theta_{10}^{y_{10}}\theta_{01}^{y_{01}}\theta_{00}^{y_{00}}\theta_{1+}^{z_{1+}}\theta_{0+}^{z_{0+}}\theta_{+1}^{u_{+1}}\theta_{+0}^{u_{+0}}. \tag{1}$$

With no missing data, i.e. the observed data are $D_{\text{c}}=(Y_{11},Y_{10},Y_{01},Y_{00})$, Altham (1971) developed Bayesian one-sided p-values for testing the null hypothesis $H_0:\theta_{1+}=\theta_{+1}$ *versus* the two one-sided alternatives. In particular, suppose that the joint prior density of $\theta=(\theta_{11},\theta_{10},\theta_{01},\theta_{00})$ is Dirichlet with parameters α_{jk} given by

$$\pi(\theta)\propto\theta_{11}^{\alpha_{11}-1}\theta_{10}^{\alpha_{10}-1}\theta_{01}^{\alpha_{01}-1}\theta_{00}^{\alpha_{00}-1}. \tag{2}$$

With no missing data the joint posterior distribution of θ is then proportional to

$$p(\theta|D_{\text{c}})\propto\theta_{11}^{y_{11}+\alpha_{11}-1}\theta_{10}^{y_{10}+\alpha_{10}-1}\theta_{01}^{y_{01}+\alpha_{01}-1}\theta_{00}^{y_{00}+\alpha_{00}-1}.$$

Table 2. Notation for the 2 × 2 table with missing data

Row variable	*Column variable*			
	1	*0*	*Missing*	*Total*
1	Y_{11}	Y_{10}	Z_{1+}	$Y_{1+}+Z_{1+}$
0	Y_{01}	Y_{00}	Z_{0+}	$Y_{0+}+Z_{0+}$
Missing	U_{+1}	U_{+0}	0	U_{++}
Total	$Y_{+1}+U_{+1}$	$Y_{+0}+U_{+0}$	Z_{++}	n

Using this posterior distribution, Altham's (1971) proposed p-value for the one-sided alternative $H_{\mathrm{A}}: \theta_{1+} < \theta_{+1}$ is the tail probability

$$p = \mathrm{pr}(\theta_{1+} > \theta_{+1} | D_{\mathrm{c}}) = \mathrm{pr}\{\theta_{10}/(\theta_{10} + \theta_{01}) > \tfrac{1}{2} | D_{\mathrm{c}}\}. \tag{3}$$

Similarly, for $H_{\mathrm{A}}: \theta_{1+} > \theta_{+1}$, Altham proposed

$$p = \mathrm{pr}(\theta_{1+} < \theta_{+1} | D_{\mathrm{c}}) = \mathrm{pr}\{\theta_{10}/(\theta_{10} + \theta_{01}) < \tfrac{1}{2} | D_{\mathrm{c}}\}.$$

For the improper Dirichlet prior with $\alpha_{10} = 1$ and $\alpha_{01} = 0$ (and $\alpha_{11} = \alpha_{00} = 0$) in expression (2), Altham (1971) showed the posterior probability that $\theta_{1+} > \theta_{+1}$ is identical to the exact McNemar p-value (Mosteller, 1952) for the one-sided alternative $H_{\mathrm{A}}: \theta_{1+} < \theta_{+1}$, namely

$$p = \sum_{r=0}^{y_{10}} \binom{y_{10} + y_{01}}{r} \times 0.5^{y_{10} + y_{01}}. \tag{4}$$

Similarly, for the improper Dirichlet prior with $\alpha_{10} = 0$ and $\alpha_{01} = 1$ (and $\alpha_{11} = \alpha_{00} = 0$) in expression (2), Altham (1971) showed that the posterior probability that $\theta_{1+} < \theta_{+1}$ is identical to the exact McNemar p-value for the one-sided alternative $H_{\mathrm{A}}: \theta_{1+} > \theta_{+1}$. Thus, Altham's Bayesian p-values with these priors are identical to the two exact one-sided McNemar p-values; we refer to a p-value that is obtained from using one of these improper Dirichlet priors as a p-value using 'the McNemar prior'.

Here, we extend the approach of Altham to formulate Bayesian p-values by calculating the posterior probability that $\theta_{1+} > \theta_{+1}$ and the posterior probability that $\theta_{1+} < \theta_{+1}$ for 2 × 2 tables with missing data. Using a Dirichlet prior as in expression (2) and assuming an ignorable missing data mechanism for the data D in Table 2 so that the likelihood is given by expression (1), the joint posterior distribution of θ is

$$p(\theta|D) \propto \theta_{11}^{y_{11} + \alpha_{11} - 1} \theta_{10}^{y_{10} + \alpha_{10} - 1} \theta_{01}^{y_{01} + \alpha_{01} - 1} \theta_{00}^{y_{00} + \alpha_{00} - 1} \theta_{1+}^{z_{1+}} \theta_{0+}^{z_{0+}} \theta_{+1}^{u_{+1}} \theta_{+0}^{u_{+0}}. \tag{5}$$

The Bayesian posterior distribution in expression (5) always exists, although it may not be unimodal. Using this posterior distribution, our proposed p-value for the one-sided alternative $H_{\mathrm{A}}: \theta_{1+} < \theta_{+1}$ is similar to equation (3), but is based on all data D,

$$p = \mathrm{pr}(\theta_{1+} > \theta_{+1} | D) = \mathrm{pr}\{\theta_{10}/(\theta_{10} + \theta_{01}) > \tfrac{1}{2} | D\}. \tag{6}$$

Similarly, the p-value for the one-sided alternative $H_{\mathrm{A}}: \theta_{1+} > \theta_{+1}$ is the posterior probability

$$p = \mathrm{pr}(\theta_{1+} < \theta_{+1} | D) = \mathrm{pr}\{\theta_{10}/(\theta_{10} + \theta_{01}) < \tfrac{1}{2} | D\}.$$

We shall explore these Bayesian p-values by using the values of (α_{11}, α_{10}, α_{01}, α_{00}) which correspond to the improper McNemar Dirichlet prior, so that the results reduce to the exact McNemar results with no missing data. Further, in the example and simulations in the following sections, we also consider the priors $\alpha_{jk} = \alpha = 0.5$ for all j and k; $\alpha_{jk} = \alpha = 1.0$ for all j and k, and $\alpha_{jk} = \alpha = 1.5$ for all j and k. To give some intuition about these other priors, if there are no missing data, and estimation of the usual odds ratio in a 2 × 2 table is of interest, $\alpha = 0.5$ (Jeffreys non-informative prior) gives a posterior mean with minimum mean-square error (Santner and Duffy, 1989), $\alpha = 1.0$ gives a posterior mode which is the usual MLE and $\alpha = 1.5$ gives a posterior mode which is the usual MLE after adding 0.5 to each cell count in the 2 × 2 table.

Unfortunately, in the presence of missing data, for any specification of (α_{11}, α_{10}, α_{01}, α_{00}), the one-sided probability in equation (6) does not have a simple closed form expression like equation (4). Since the dimension of the integration is low (three dimensions), the posterior probabilities in expression (6) can be computed by using direct numerical integration; we used quasi-Monte-Carlo integration (Niederreiter, 1978; González *et al.*, 2006) with 65000 quadrature points for the example and simulations. An SAS macro (SAS Institute, 2004) is available from the first author. For any sample size, the SAS macro takes approximately 2 s to calculate a p-value on a Pentium 4, 2.4-GHz, 1-Gbyte random-access memory computer. Alternatively, Markov chain Monte Carlo sampling via the software WinBUGS (Spiegelhalter *et al.*, 2004) can be used.

Altham did not discuss the two-sided alternative $H_A : \theta_{1+} \neq \theta_{+1}$. For this two-sided alternative, the posterior probability of the null hypothesis is a point probability and thus not appropriate to use. With no missing data, the exact distribution of McNemar's test statistic is symmetric under the null hypothesis, so the exact two-sided p-value is two times the minimum of the two one-sided p-values. Here, for our Bayesian method, we also propose to calculate a two-sided p-value as two times the minimum of the two one-sided p-values

$$p = 2 \min\left[\mathrm{pr}\{\theta_{10}/(\theta_{10}+\theta_{01}) > \tfrac{1}{2}|D\}, \mathrm{pr}\{\theta_{10}/(\theta_{10}+\theta_{01}) < \tfrac{1}{2}|D\}\right]. \tag{7}$$

Fleiss *et al.* (2003) showed that a two-sided p-value that is calculated as two times the minimum of two one-sided p-values will have the correct type I error.

4. Analysis of data application

In this section, we illustrate the key results of this paper by using data from the study of malpractice cases (Greenberg *et al.*, 2007) that was discussed in Section 2. 69 communication breakdowns were reviewed by two surgeon–reviewers, who were asked to determine whether the communication breakdown occurred during a hand-off in care (with possible answer 'yes' or 'no'). The data are given in Table 1.

Table 3 shows the one- and two-sided p-values for the various priors of interest, using all the data as well as only complete cases. We present the one-sided p-values for completeness, but there was no predetermined alternative that one rater was more likely to find problems with a hand-off of care than the other, so the alternative of interest is the two-sided alternative. In Table 3, we see that a p-value that is based on only the complete cases is approximately twice the size of the corresponding p-values that are based on all data for any given prior. We also see that, for both complete cases and all data, the McNemar prior gives the most conservative

Table 3. p-values for the example with a null hypothesis of the equal probabilities of two raters' finding problems with hand-off of care

Alternative hypothesis	*Results for the following priors:*							
	McNemar prior		$\alpha = 0.5$		$\alpha = 1.0$		$\alpha = 1.5$	
	All data	*Complete cases*	*All data*	*Complete cases*	*All data*	*Complete cases*	*All data*	*Complete cases*
$H_A : \theta_{10} \neq \theta_{01}$	0.120	0.219	0.048	0.090	0.064	0.121	0.079	0.149
$H_A : \theta_{10} < \theta_{01}$	0.060	0.109	0.024	0.045	0.032	0.060	0.040	0.074
$H_A : \theta_{10} > \theta_{01}$	0.992	0.984	0.976	0.954	0.968	0.940	0.960	0.929

p-value among all the priors; except for the McNemar prior, the p-values by using all the data are significant at the 10% level, whereas the p-values by using complete cases are not. Finally, using the prior $\alpha = 0.5$ gives the smallest p-values; only the two-sided p-value by using the prior $\alpha = 0.5$ with all the data was significant at a 5% level of significance.

This example highlights how different priors and/or complete cases *versus* all data can produce different results. However, this is just one data set; to put the results of this data set in the context of possible missingness at random (MAR) mechanisms in 2 × 2 tables, as well as to examine the finite sample properties of the p-values by using different priors and complete cases or all the data, we conduct a simulation study in the following section.

5. Simulation study

In the simulation study, we consider a scenario in which the cell probabilities $\theta_{11} = 0.25$ and $\theta_{00} = 0.25$; θ_{10} varies from 0.25 to 0.45 in increments of 0.025 and, correspondingly, θ_{01} decreases from 0.25 to 0.05 in increments of 0.025. The exact McNemar test statistic is based on the conditional probability $\theta_{10}/(\theta_{10} + \theta_{01})$, which equals 0.5 under $H_0 : \theta_{10} = \theta_{01}$. For these simulations, the conditional probability $\theta_{10}/(\theta_{10} + \theta_{01})$ ranges from the null hypothesis of 0.5–0.9 in increments of 0.05. We chose total sample sizes of $n = 60$ and $n = 250$ and fixed the number of simulation replications at 2000 for each value of $\theta = (\theta_{11}, \theta_{10}, \theta_{01}, \theta_{00})$. Note that, for a given value θ, each of the 2000 replications can have different proportions of missing data.

In the simulations, we explore the rejection percentages (type I error rate and power) for a 5% significance level test using the two-sided alternative $H_A : \theta_{1+} \neq \theta_{+1}$, with p-value calculated as two times the minimum of the one-sided p-values. For a given set of 2000 simulation replications, the rejection percentage is calculated as the percentage of times that the p-value is less than 0.05; the standard error of the estimated rejection percentage is 1.1% when the true rejection percentage is 50% and 0.5% when the true rejection percentage is 5%. We calculate the p-values by using all the data and just the complete cases; we use the McNemar prior, and the priors $\alpha = 0.5$, $\alpha = 1.0$ and $\alpha = 1.5$. Further, we specified two possible MAR missing data mechanisms (Chen and Fienberg, 1974), as we now discuss.

To describe the MAR mechanism of Chen and Fienberg (1974), we let V_1 denote the row variable and V_2 denote the column variable, with $V_m = 0, 1$ for $m = 1, 2$. Further, we define the missing data indicator random variable R_m = 1 if V_m is missing and 0 if V_m is observed, for $m = 1, 2$. The 'missing data mechanism' is the conditional multinomial distribution of the missing data indicators (R_1, R_2) given (V_1, V_2), with probabilities $\Pr(R_1 = r_1, R_2 = r_2 | V_1 = v_1, V_2 = v_2)$. In this MAR model, the probability that both V_1 and V_2 are missing is 0, i.e. $\Pr(R_1 = 1, R_2 = 1 | V_1 = v_1, V_2 = v_2) = 0$. The probability that V_1 is observed and V_2 is missing depends only on the observed data V_1,

$$\Pr(R_1 = 0, R_2 = 1 | V_1 = v_1, V_2 = v_2) = \varphi_{v1} = v_1\varphi_1 + (1 - v_1)\varphi_0, \tag{8}$$

where the probabilities (φ_1, φ_0) must be specified. Similarly, the probability that V_2 is observed and V_1 is missing depends only on the observed data V_2,

$$\Pr(R_1 = 1, R_2 = 0 | V_1 = v_1, V_2 = v_2) = \rho_{v2} = v_2\rho_1 + (1 - v_2)\rho_0, \tag{9}$$

where the probabilities (ρ_1, ρ_0) must be specified. Since the probability that both V_1 and V_2 are missing is 0, the probability that both V_1 and V_2 are observed is $1 - \varphi_{v1} - \rho_{v2}$, which depends on the values of both V_1 and V_2. This MAR missing data mechanism satisfies the assumption that the probability that a row or column variable is missing depends only on observed data. A given missing data configuration entails specification of the four probabilities $(\varphi_1, \varphi_0, \rho_1, \rho_0)$. In the simulations, we specified two sets of $(\varphi_1, \varphi_0, \rho_1, \rho_0)$: the first set (MAR1) is $(\varphi_1, \varphi_0, \rho_1, \rho_0) = (0.3, 0.15, 0.15, 0.4)$, and the second set (MAR2) is $(\varphi_1, \varphi_0, \rho_1, \rho_0) = (0.3, 0.15, 0.3, 0.15)$.

Before discussing the results of the simulations, we give insight into why the type I error can be elevated under MAR when restricting the analysis to 'complete cases', i.e. subjects with $(R_1 = 1,\ R_2 = 1)$. With no missing data, the exact McNemar test statistic tests that the conditional probability of being in cell (1,0) given in cell (1,0) or (0,1), e.g. $\theta_{10}/(\theta_{10} + \theta_{01})$, equals 0.5 under $H_0 : \theta_{10} = \theta_{01}$. Suppose that, in complete cases, we perform an exact McNemar test that the conditional probability of being in cell (1,0) given in cell (1,0) or (0,1) equals 0.5. Using Bayes's rule with equations (8) and (9), in complete cases, the conditional probability of being in cell (1,0) given in cell (1,0) or (0,1) equals

$$\begin{aligned}\pi &= \Pr(V_1 = 1, V_2 = 0 | R_1 = 1, R_2 = 1, V_1 + V_2 = 1)\\ &= \frac{\theta_{10}(1 - \varphi_1 - \rho_0)}{\theta_{10}(1 - \varphi_1 - \rho_0) + \theta_{01}(1 - \varphi_0 - \rho_1)}.\end{aligned} \tag{10}$$

Then, with complete cases, when testing that $\pi = 0.5$, we are not necessarily testing the null hypothesis that $\theta_{10}/(\theta_{10} + \theta_{01}) = 0.5$. For the MAR mechanism that was used in the simulation, under $H_0 : \theta_{10} = \theta_{01}$,

Table 4. Simulation rejection percentages for two-sided Bayesian *p*-values for MAR model 1 for a 5% significance level test for $H_0 : \theta_{10} = \theta_{01}$, as the conditional probability $\theta_{10}/(\theta_{10} + \theta_{01})$ increases from the null value of 0.50

$\frac{\theta_{10}}{\theta_{10}+\theta_{01}}$	*Results for the following priors:*							
	McNemar prior		$\alpha = 0.5$		$\alpha = 1.0$		$\alpha = 1.5$	
	All data	*Complete cases*	*All data*	*Complete cases*	*All data*	*Complete cases*	*All data*	*Complete cases*
n = 60								
0.50 (null)	2.55	24.30	4.85	34.50	4.30	31.75	3.55	28.95
0.55	6.20	14.00	9.00	20.60	8.00	17.8	7.60	16.00
0.60	10.85	7.95	14.50	13.90	13.45	12.35	12.30	10.45
0.65	22.60	4.35	27.95	7.55	25.95	6.55	24.90	5.75
0.70	40.85	2.80	46.95	5.75	45.45	4.50	44.25	4.00
0.75	58.40	3.55	65.45	6.85	64.05	5.95	62.10	4.95
0.80	77.55	9.45	82.70	15.25	81.80	13.30	80.95	12.15
0.85	89.65	17.45	93.05	26.50	92.75	24.20	92.25	21.65
0.90	97.00	35.80	98.10	47.85	98.00	45.45	97.85	40.60
n = 250								
0.50 (null)	3.75	86.55	5.35	89.25	5.10	89.00	4.95	88.60
0.55	13.60	62.80	14.90	68.60	14.70	67.90	14.60	67.10
0.60	45.50	31.20	48.60	36.90	48.40	36.00	48.10	35.25
0.65	79.25	9.95	81.80	11.85	81.70	11.65	81.50	11.45
0.70	96.30	3.95	97.05	5.35	97.00	5.25	97.00	5.00
0.75	99.60	11.10	99.80	13.15	99.80	13.10	99.80	12.65
0.80	100.00	38.80	100.00	43.05	100.00	42.85	100.00	41.55
0.85	100.00	78.15	100.00	82.00	100.00	81.80	100.00	80.70
0.90	100.00	97.50	100.00	98.40	100.00	98.30	100.00	98.20

$$\pi = \frac{1-\varphi_1-\rho_0}{1-\varphi_1-\rho_0+1-\varphi_0-\rho_1},$$

which may not equal 0.5. The further π is from 0.5 under $H_0:\theta_{10}=\theta_{01}$, the larger the type I error when using only complete cases. For set MAR1, with $(\varphi_1,\varphi_0,\rho_1,\rho_0)=(0.3,0.15,0.15,0.4)$, we have $\pi=0.3$ under $H_0:\theta_{10}=\theta_0$; we expect that a test that is based on complete cases will have a much higher type I error than the nominal 5%. For set MAR2, with $(\varphi_1,\varphi_0,\rho_1,\rho_0)=(0.3,0.15,0.3,0.15)$, we have $\pi=\theta_{10}/(\theta_{10}+\theta_{01})$ $(=0.5$ under $H_0:\theta_{10}=\theta_{01})$, and we expect that a test which is based on complete cases will have the specified type I error of 5% but may be less powerful than the test which is based on all the data.

Table 4 shows the rejection percentages for set MAR1. As discussed above, the complete cases could give a very high type 1 error since the complete-case null value of π does not equal 0.5, but equals 0.3. As we see in Table 4, this is so, in that the type I error is greater than 20% for all complete-case tests, whereas the type I error is approximately 5%, using 'all data', Further, we see that the rejection percentages (estimated power) for complete cases actually decreases as $\theta_{10}/(\theta_{10}+\theta_{01})$ increases from 0.5 to 0.7, mainly because π in equation (10) accordingly increases from 0.3 to 0.5, so that the 'complete-case null value' $\pi=0.5$ actually occurs when $\theta_{10}/(\theta_{10}+\theta_{01})=0.7$. Finally, using 'all data' for $n=60$ and $n=250$, the McNemar prior gives the most conservative test, and the prior $\alpha=0.5$ gives the most powerful test, although the power becomes similar for all priors when $n=250$.

Using set MAR2 (the results are not shown), for both $n=60$ and $n=250$, all tests appear to have type 1 error no more than 5%. Again, the McNemar prior using both all the data and complete cases is the most conservative, and the Bayesian tests using all the data are much more powerful (as much as 15% in absolute terms), with the test with prior $\alpha=0.5$ having the highest power.

6. Conclusion

In this paper, we explore Bayesian p-values in 2×2 tables from matched pairs with missing data that are missing at random. These tests are particularly useful when the cell sizes are small, since large sample test statistics with missing data such as the likelihood ratio statistic (Gokhale and Sirtonik, 1984) are not always available because there may be no unique MLE under the alternative and/or the null hypothesis. In simulating data that are missing at random, we found that the Bayesian p-values by using all the data have type I error rates that are close to the nominal level, as compared with complete-case p-values, which can have a type I error rate that is much higher than the nominal level. Even when the complete-case analysis gives the correct type I error, use of all the data with our Bayesian approach increases the power.

The p-values for the example in Table 3 seem to be consistent with the simulations in Table 4. In particular, in the example, when using all of the data, all priors (except the McNemar prior) give p-values that are significant at the 10% level, whereas none of the complete-case p-values are significant. Further, the p-values by using complete cases are approximately twice the size of the p-values by using all the data. Thus, the example may be similar to a data set from the first MAR mechanism in the simulations, in which the alternative may be true, and we reject the null hypothesis using all of the data, but we do not reject using only complete cases. The example is also consistent with the simulations in that the McNemar prior gives the most conservative result.

Because of the broad range of possible missing data mechanisms, it is difficult to draw definitive conclusions from a simulation study. Nonetheless, in terms of type I error and power, in the simulation study that is reported here, the Bayesian p-values appear to perform discernibly better than the p-values that are based on complete cases. On the basis of our results, we would not use a complete-case analysis.

The Bayesian approach that is taken in this paper can also be used for related problems with missing data in matched pairs studies. For example, suppose that we are interested in estimation instead of testing; the posterior means from the Bayesian posterior distributions that are discussed in this paper could be used for estimating parameters of interest, such as the difference in marginal probabilities $(\theta_{10} - \theta_{01})$ or the odds θ_{10}/θ_{01}. Further, suppose that missingness depends on other observed variables besides the row and/or column variables, such as subject and reviewer characteristics; a Bayesian approach which incorporates these other variables in the likelihood could be developed. These topics warrant further exploration.

Acknowledgements

We are grateful for the support from the US National Institutes of Health grants AI 60373, MH 054693, CA 74015 and CA 069222. We are also extremely indebted to Dr David Studdert for allowing the data to be used in this analysis.

References

Altham, P. M. E. (1969) Exact Bayesian analysis of a 2 × 2 contingency table, and Fisher's "exact" significance test. *J. R. Statist. Soc.* B, **31**, 261–269.

Altham, P. M. E. (1971) The analysis of matched proportions. *Biometrika*, **58**, 561–576.

Berger, J. O. and Delampady, M. (1987) Testing precise hypotheses. *Statist. Sci.*, **2**, 317–352.

Berger, J. O. and Sellke, T. (1987) Testing a point null hypothesis: the irreconcilability of *p*-values and evidence (with discussion). *J. Am. Statist. Ass.*, **82**, 112–122.

Casella, G. and Berger, R. L. (1987) Reconciling Bayesian and frequentist evidence in the one-sided testing problem (with discussion). *J. Am. Statist. Ass.*, **82**, 106–111.

Chen, H. Y. and Little, R. J. A. (1999) A test of missing completely at random for generalized estimating equations with missing data. *Biometrika*, **86**, 1–13.

Chen, T. and Fienberg, S. E. (1974) Two-dimensional contingency tables with both completely and partially cross-classified data. *Biometrics*, **30**, 629–642.

Fleiss, J. L., Levin, B. and Paik, M. C. (2003) *Statistical Methods for Rates and Proportions*, 3rd edn. New York: Wiley.

Fuchs, C. (1982), Maximum likelihood estimation and model selection in contingency tables with missing data. *J. Am. Statist. Ass.*, **77**, 270–278.

Gokhale, D. and Sirtonik, B. (1984) On tests for correlated proportions in the presence of incomplete data. *Psychometrika*, **49**, 147–152.

González, J., Tuerlinckx, F., De Boeck, P. and Cools, R. (2006) Numerical integration in logistic-normal models. *Computnl Statist. Data Anal.*, **51**, 1535–1548.

Greenberg, C. C., Regenbogen, S. E., Studdert, D. M., Lipsitz, S. R., Rogers, S. O., Zinner, M. J. and Gawande, A. A. (2007) Patterns of communication breakdowns resulting in injury to surgical patients. *J. Am. Coll. Surg.*, **204**, 533–540.

McNemar, Q. (1947) Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, **12**, 153–157.

Mosteller, F. (1952) Some statistical problems in measuring the subjective response to drugs. *Biometrics*, **8**, 220–226.

Niederreiter, H. (1978) Quasi-Monte Carlo methods and pseudo-random numbers. *Bull. Am. Math. Soc.*, **84**, 957–1041.

Rogers, S. O., Gawande, A. A., Kwaan, M., Puopolo, A. L., Yoon, C., Brennan, T. A. and Studdert, D. M. (2006) Analysis of surgical errors in closed malpractice claims at 4 liability insurers. *Surgery*, **140**, 25–33.

Rubin, D. B. (1976) Inference and missing data. *Biometrika*, **63**, 581–592.

Santner, T. J. and Duffy, D. E. (1989) *The Statistical Analysis of Discrete Data*. New York: Springer.

SAS Institute (2004) *SAS® 9.1 Macro Language: Reference*. Cary: SAS Institute.

Spiegelhalter, D., Thomas, A., Best, N. and Lunn, D. (2004) *WinBUGS User Manual Version 2.0*. Cambridge: Medical Research Council Biostatistics Unit. (Available from `http://mathstat.helsinki.fi/openbugs/data/Docu/WinBugs%20Manual.html`.)