Computational Statistics and Data Analysis 127 (2018) 69–81

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Assessing non-inferiority for incomplete paired-data under non-ignorable missing mechanism

Huiqiong Li [a], Guoliang Tian [b], Niansheng Tang [a], Hongyuan Cao [c,d,*]

[a] Department of Statistics, Yunnan University, PR China
[b] Department of Mathematics, Southern University of Science and Technology, PR China
[c] Center for Applied Statistical Research, School of Mathematics, Jilin University, PR China
[d] Department of Statistics, University of Missouri-Columbia, USA

ARTICLE INFO

Article history:
Received 20 December 2016
Received in revised form 29 April 2018
Accepted 12 May 2018
Available online 26 May 2018

Keywords:
Bayesian *p*-values
Credible intervals
Highest posterior density intervals
Non-inferiority test
Incomplete paired data

ABSTRACT

Testing equivalence of incomplete paired data arises frequently in biomedical studies. Most existing work impose the missing at random assumption, which is not realistic in practice. Two Bayesian approaches for testing the non-inferiority of incomplete paired data under non-ignorable missing mechanism are presented. In addition, Bayesian credible intervals and highest posterior density intervals for the risk difference are constructed. Simulation studies are conducted to evaluate the performance of the two Bayesian testing procedures and the credible intervals. Two datasets are used to illustrate the proposed methods.

1. Introduction

Assessing the non-inferiority of a new method or treatment with the standard one is an important topic in comparative clinical studies. Non-inferiority trials are often employed to evaluate whether a less toxic, easier to administer or inexpensive new treatment is not inferior to the standard treatment in terms of efficacy. Non-inferiority assessment has received a lot of attention for matched-pair trials in the past decades. For example, Tango (1998) derived a score statistic to test non-inferiority via relative risk in a re-parameterized model with a matched-pair design. Tang et al. (2003) developed an alternative score test procedure to test equivalence or non-inferiority via relative risk in a matched-pair design. Chan et al. (2003) proposed an exact method to assess non-inferiority via rate ratio with small-sample matched-pair design.

In practice, in comparative studies of two treatments or reviewers, incomplete matched-pair data are often encountered. For example, in a study of medical malpractice cases (Greenberg et al., 2007; Lin et al., 2009; Altham and Hankin, 2010; Konietschke et al., 2012), two surgeon-reviewers used a structured instrument to evaluate 69 errors, and to identify important human and system factors contributing to the errors. Among many possible factors is communication breakdown, each surgeon-reviewer was asked to determine whether a handoff in care was associated with the communication breakdown. In this study, 8 reviews were missing for Surgeon 1 and 11 reviews were missing for Surgeon 2. Thus, the resultant data include two parts: the complete observations and the incomplete observations. This dataset is displayed in Table 1.

Under the assumption of missing at random (MAR), the probability of missing only depends on observed data. In the case of MAR, various authors have studied the problem of the equivalence test and confidence interval construction for

* Corresponding author at: Department of Statistics, University of Missouri-Columbia, USA.
E-mail address: caohong@missouri.edu (H. Cao).

https://doi.org/10.1016/j.csda.2018.05.009
0167-9473/

Table 1
Counts of two surgeon reviewers' answer in a study of medical malpractice.

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

Table 2
Data structure for a matched-pair design with missing observations.

Treatment 1	Treatment 2		
	Positive response ($Y = 1$)	Negative response ($Y = 0$)	Missing
Positive response ($X = 1$)	$n_1(\theta_1)$	$n_2(\theta_2)$	m_{12}
Negative response ($X = 0$)	$n_3(\theta_3)$	$n_4(\theta_4)$	m_{34}
Missing	m_{13}	m_{24}	m_{1234}

two correlated proportions with incomplete matched-pair data (Choi and Stablein, 1982; Ekbohm, 1982; Tang and Tang, 2004; Tang et al., 2009; Lin et al., 2009; Tang et al., 2011). Non-ignorable missing or missing not at random (MNAR) refers to the case that the probability of missing is related to the value of the missing data. In the case of MNAR, Choi and Stablein (1988) proposed several methods for testing the equality of two correlated proportions. Nandram and Choi (2002) proposed a Bayesian approach for a non-ignorable non-response model. To the best of our knowledge, there is no published work to date that deals with incomplete paired-data under non-ignorable missing mechanism. In this paper, we develop Bayesian methods to test non-inferiority and to construct Bayesian credible intervals and highest posterior density (HPD) intervals for incomplete paired-data.

The rest of this paper is organized as follows. In Section 2, we present two Bayesian p-values to assess non-inferiority for incomplete paired data under the non-ignorable missing mechanism. Section 3 develops a new Bayesian interval estimation of the risk difference for incomplete paired data under nonignorable missing mechanism. Simulation studies are conducted to investigate the performance of various methods in Section 4. We illustrate the proposed methodology with two dataset in Section 5. Concluding remarks are given in Section 6.

2. Bayesian methods for the non-inferiority test

2.1. Data structure and non-inferiority test

Consider a trial for comparing two treatments. Suppose that X and Y are two correlated binary variables. Let $X = 1$ (or $X = 0$) if a subject has a positive (or negative) response under treatment 1 and let $Y = 1$ (or $Y = 0$) if the same subject has a positive (or negative) response under treatment 2. Let $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_4)^\top$ denote model parameters, where $\theta_1 = \Pr(X = 1, Y = 1)$, $\theta_2 = \Pr(X = 1, Y = 0)$, $\theta_3 = \Pr(X = 0, Y = 1)$ and $\theta_4 = \Pr(X = 0, Y = 0)$. Naturally, we have $\boldsymbol{\theta} \in \mathbb{T}_4$, where $\mathbb{T}_n := \{(x_1, \ldots, x_n)^\top : x_i > 0, \sum_{i=1}^n x_i = 1\}$.

Suppose that in a comparative trial there are a total of N participants containing $n = \sum_{j=1}^4 n_j$ complete cases and $m_{12} + m_{34} + m_{13} + m_{24} + m_{1234}$ incomplete cases, where n_1 subjects have both positive responses, n_2 subjects have a positive response for Treatment 1 and a negative response for Treatment 2, n_3 subjects have a negative response for Treatment 1 and a positive response for Treatment 2, n_4 subjects have both negative responses; m_{12} (or m_{34}) subjects only have a positive (or negative) response for Treatment 1, m_{13} (or m_{24}) subjects only have a positive (or negative) response for Treatment 2; and the responses for m_{1234} subjects are totally missing for both treatments. These observed outcomes are reported in Table 2. We denote the observed data by $Y_{\text{obs}} = \{n_1, \ldots, n_4; m_{12}, m_{34}, m_{13}, m_{24}, m_{1234}\}$ with $N = \sum_{j=1}^4 n_j + m_{12} + m_{34} + m_{13} + m_{24} + m_{1234}$.

Treatment 1 is said to be not inferior to Treatment 2 if $\Pr(X = 1) > \Pr(Y = 1) - \delta_0$, i.e., $\theta_1 + \theta_2 > \theta_1 + \theta_3 - \delta_0$, where $\delta_0 > 0$ is the non-inferiority margin of clinical interest. Thus, testing the non-inferiority of Treatment 1 to Treatment 2 is equivalent to testing the following hypothesis:

$$H_0: \theta_2 \leqslant \theta_3 - \delta_0 \quad \text{against} \quad H_1: \theta_2 > \theta_3 - \delta_0. \tag{2.1}$$

The objective of this paper is to develop Bayesian methods for testing H_0 versus H_1 under the non-ignorable missing mechanism.

2.2. Formulation of the non-ignorable missing mechanism

To describe the non-ignorable missing mechanism in Table 2, we first define a 4-category response random variable R, where $R = 12$ if a subject has response to both treatments, $R = 1\bar{2}$ if a subject has response only to Treatment 1, $R = \bar{1}2$

Table 3
The observed counts, the marginal probabilities of Z and R, and the joint distribution of (Z, R).

$Z = (X, Y)$	Response variable R						Marginal probability of Z
	12	$1\bar{2}$	$\bar{1}2$	$\bar{1}2$	$\bar{1}2$	$\bar{1}\bar{2}$	
$(1, 1)$	$\pi_{11}\,(n_1)$	π_{21}	0	π_{31}	0	π_{41}	θ_1
$(1, 0)$	$\pi_{12}\,(n_2)$	π_{22}	0	0	π_{32}	π_{42}	θ_2
$(0, 1)$	$\pi_{13}\,(n_3)$	0	π_{23}	π_{33}	0	π_{43}	θ_3
$(0, 0)$	$\pi_{14}\,(n_4)$	0	π_{24}	0	π_{34}	π_{44}	θ_4
Marginal probability of R	ϕ_1	ϕ_2		ϕ_3		ϕ_4	1
Observed counts	n	m_{12}	m_{34}	m_{13}	m_{24}	m_{1234}	N

if a subject has response only to Treatment 2 and $R = \bar{1}\bar{2}$ if a subject has no response to both treatments. The probability mass function of R is represented by $\phi_1 = \Pr(R = 12)$, $\phi_2 = \Pr(R = 1\bar{2})$, $\phi_3 = \Pr(R = \bar{1}2)$ and $\phi_4 = \Pr(R = \bar{1}\bar{2})$. Let $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_4)^\top$. Then $\boldsymbol{\phi}$ is called the parameter vector associated with the missing data mechanism.

For convenience, define $Z = (X, Y)$. The joint distribution of Z and R is denoted by $\boldsymbol{\pi} = (\pi_{ij})$ for $i, j = 1, \ldots, 4$. For example, $\pi_{11} = \Pr\{Z = (1, 1), R = 12\}$ and $\pi_{44} = \Pr\{Z = (0, 0), R = \bar{1}\bar{2}\}$. Table 3 shows the observed counts, the marginal probabilities of Z and R, and the joint distribution of (Z, R). The likelihood function for $\boldsymbol{\pi}$ based on the observed data D_{obs} is proportional to (Tian et al., 2003)

$$\left(\prod_{j=1}^{4}\pi_{1j}^{n_j}\right)(\pi_{21}+\pi_{22})^{m_{12}}(\pi_{23}+\pi_{24})^{m_{34}}(\pi_{31}+\pi_{33})^{m_{13}}(\pi_{32}+\pi_{34})^{m_{24}}\left(\sum_{j=1}^{4}\pi_{4j}\right)^{m_{1234}}. \tag{2.2}$$

If Z and R are independent,

$$\pi_{ij} = \phi_i\theta_j, \quad i, j = 1, \ldots, 4,$$

the missing data mechanism is ignorable or MAR (Little and Rubin, 2002). As $\sum_{i=1}^{4}\theta_i = 1$ and $\sum_{i=1}^{4}\phi_i = 1$, under MAR, there are only 6 free parameters and (2.2) becomes

$$\left(\prod_{j=1}^{4}\theta_j^{n_j}\right)(\theta_1+\theta_2)^{m_{12}}(\theta_3+\theta_4)^{m_{34}}(\theta_1+\theta_3)^{m_{13}}(\theta_2+\theta_4)^{m_{24}}$$
$$\times\ \phi_1^{n}\phi_2^{m_{12}+m_{34}}\phi_3^{m_{13}+m_{24}}\phi_4^{m_{1234}}, \quad \boldsymbol{\theta}\in\mathbb{T}_4, \quad \boldsymbol{\phi}\in\mathbb{T}_4. \tag{2.3}$$

Based on (2.3), we can obtain the maximum likelihood estimates (MLEs) of $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ with explicit expressions. In other words, both $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ are estimable in the frequentist framework.

If Z and R are dependent, we call the missing data mechanism non-ignorable. Under the non-ignorable missing mechanism, we have 15 free parameters since $\boldsymbol{\pi} \in \mathbb{T}_{16}$. These $\{\pi_{ij}\}$ are not estimable in the frequentist framework. Yet, they are identifiable if we put a prior distribution on π.

2.3. Bayesian tests

The Bayes factor is a very useful tool to test hypothesis. The Bayes factor for comparing H_1 to H_0 is given by

$$BF_{10} = \frac{m_1(D_{obs})}{m_0(D_{obs})}, \tag{2.4}$$

where

$$m_j(D_{obs}) = \int p(D_{obs}|\boldsymbol{\pi}_j)f(\boldsymbol{\pi}_j)d\boldsymbol{\pi}_j, \tag{2.5}$$

where $p(D_{obs}|\boldsymbol{\pi}_j)$ is the likelihood function of the observed data and $f(\boldsymbol{\pi}_j)$ is the prior for the unknown parameter under H_j, $j = 0, 1$, respectively.

The Bayes factor provides evidence in favor of H_1 against H_0. According to Kass and Raftery (1995), if $0 \leqslant \log(\mathrm{BF}_{10}) < 1$, the evidence against H_0 is "not worth more than a bare mention"; if $1 \leqslant \log(\mathrm{BF}_{10}) < 3$, the evidence against H_0 is "positive"; if $3 \leqslant \log(\mathrm{BF}_{10}) < 5$, the evidence against H_0 is "strong"; and if $\log(\mathrm{BF}_{10}) \geqslant 5$, the evidence against H_0 is "very strong".

However, the function $m_j(D_{obs})$ in (2.5) usually does not have a closed form and the integration is extremely difficult to calculate. To circumvent this, we shall develop Bayesian p-values to test H_0 against H_1 in (2.1) for incomplete paired-data under non-ignorable missing mechanism.

2.3.1. The conventional Bayesian p-value

In order to test (2.1) under non-ignorable missing mechanism, we first discuss the Bayesian p-value with complete data. Suppose the joint prior density of $(\pi_{11}, \pi_{12}, \pi_{13}, \pi_{14})$ is Dirichlet, with parameters $\alpha = (\alpha_{11}, \alpha_{12}, \alpha_{13}, \alpha_{14})$, given by the density kernel

$$f(\pi) \propto \pi_{11}^{\alpha_{11}-1}\pi_{12}^{\alpha_{12}-1}\pi_{13}^{\alpha_{13}-1}\pi_{14}^{\alpha_{14}-1}.$$

Then, if there are no missing data, the joint posterior distribution of $(\pi_{11}, \pi_{12}, \pi_{13}, \pi_{14})$ is proportional to

$$\prod_{j=1}^{4}\pi_{1j}^{n_j+\alpha_{1j}-1}.$$

Based on this posterior distribution, p-value can be obtained for the one-sided alternative $\theta_2 > \theta_3 - \delta_0$ using the tail probability (Altham, 1969, 1971)

$$p = \Pr(\theta_2 \leqslant \theta_3 - \delta_0 | D_{obs}).$$

For the improper Dirichlet prior with $\alpha_{12} = 0, \alpha_{13} = 1$ and $\alpha_{11} = \alpha_{14} = 0$, Altham (1971) showed the posterior probability that $\theta_2 \leqslant \theta_3 - \delta_0$ is identical to the exact McNemar p-value (Mosteller, 1952) for the one-sided alternative $\theta_2 > \theta_3 - \delta_0$.

Using the same Dirichlet prior, we extend the approach of Altham (1969, 1971) to develop Bayesian p-values under non-ignorable missing mechanism. By calculating the posterior probability, the proposed Bayesian p-value for the one-sided alternative $H_1 : \theta_2 > \theta_3 - \delta_0$ is the tail probability

$$p_{v1} = \Pr(\theta_2 \leqslant \theta_3 - \delta_0 | D_{obs}). \tag{2.6}$$

To compute the posterior probability under non-ignorable missing mechanism, one need to determine priors for $\{\pi_{ij}\}$. Let the joint prior for $\{\pi_{ij}\}$ be a Dirichlet distribution with known parameters $\{\alpha_{ij}\}$; that is $f(\boldsymbol{\pi}) \propto \prod_{i=1}^{4}\prod_{j=1}^{4}\pi_{ij}^{\alpha_{ij}-1}$. Then, the posterior distribution by combining the prior distribution and the likelihood function can be calculated. Through (2.2), the posterior distribution of $\boldsymbol{\pi}$ is proportional to

$$\left(\prod_{j=1}^{4}\pi_{1j}^{n_j+\alpha_{1j}-1}\right)\left(\prod_{i=2}^{4}\prod_{j=1}^{4}\pi_{ij}^{\alpha_{ij}-1}\right)(\pi_{21}+\pi_{22})^{m_{12}}(\pi_{23}+\pi_{24})^{m_{34}}$$
$$\times(\pi_{31}+\pi_{33})^{m_{13}}(\pi_{32}+\pi_{34})^{m_{24}}\left(\sum_{j=1}^{4}\pi_{4j}\right)^{m_{1234}}. \tag{2.7}$$

Unfortunately, under non-ignorable missing mechanism, for any Dirichlet prior, the Bayesian p-value in (2.6) does not have a closed form. Since the dimension of the integration is high, (2.6) can be extremely difficult to calculate by direct numerical integration. Here, we shall compute it by sampling-based method.

First, we have the following notations.

$$\begin{cases} \pi_{1i} & = & \xi_i, & 1 \leqslant i \leqslant 4, \\ \pi_{2,i-4} & = & \xi_i, & 5 \leqslant i \leqslant 8, \\ \pi_{31} & = & \xi_9, & \\ \pi_{33} & = & \xi_{10}, & \\ \pi_{32} & = & \xi_{11}, & \\ \pi_{34} & = & \xi_{12}, & \\ \pi_{4,i-12} & = & \xi_i, & 13 \leqslant i \leqslant 16. \end{cases} \tag{2.8}$$

It can be shown that the posterior distribution in (2.7) becomes

$$\left(\prod_{i=1}^{4}\xi_i^{n_i+\alpha_{1i}-1}\right)\left(\prod_{i=5}^{8}\xi_i^{\alpha_{2,i-4}-1}\right)\xi_9^{\alpha_{31}-1}\xi_{10}^{\alpha_{33}-1}\xi_{11}^{\alpha_{32}-1}\xi_{12}^{\alpha_{34}-1}\left(\prod_{i=13}^{16}\xi_i^{\alpha_{4,i-12}-1}\right)$$
$$\times\left(\sum_{j=1}^{4}\xi_j\right)^{0}\left(\sum_{j=5}^{6}\xi_j\right)^{m_{12}}\left(\sum_{j=7}^{8}\xi_j\right)^{m_{34}}\left(\sum_{j=9}^{10}\xi_j\right)^{m_{13}}\left(\sum_{j=11}^{12}\xi_j\right)^{m_{24}}\left(\sum_{j=13}^{16}\xi_j\right)^{m_{1234}}. \tag{2.9}$$

According to Tian et al. (2003), (2.9) is the kernel of a grouped Dirichlet distribution (GDD) with six partitions. The definition of GDD is given as follows.

An n-dimensional random vector $\mathbf{X} = (X_1, \ldots, X_n)^\top \in \mathbb{T}_n$ is said to follow a grouped Dirichlet distribution with m partitions if the joint distribution of $(X_1, \ldots, X_{n-1})^\top$ is given by

$$c_m^{-1} \cdot \left(\prod_{i=1}^{n} x_i^{a_i - 1} \right) \cdot \prod_{j=1}^{m} \left(\sum_{k=s_{j-1}+1}^{s_j} x_k \right)^{b_j},$$

where c_m is the normalizing constant given by

$$c_m = \left\{ \prod_{j=1}^{m} B_{s_j - s_{j-1}}(a_{s_{j-1}+1}, \ldots, a_{s_j}) \right\} \cdot B_m \left(\sum_{k=1}^{s_1} a_k + b_1, \ldots, \sum_{k=s_{m-1}+1}^{s_m} a_k + b_m \right)$$

and $B_n(a_1, \ldots, a_n) = \{\prod_{i=1}^{n} \Gamma(a_i)\} / \Gamma(\sum_{i=1}^{n} a_i)$ denotes the multivariate beta function. Write $\mathbf{X} \sim \mathrm{GD}_{n,m,\mathbf{s}}(\mathbf{a}, \mathbf{b})$ on $\mathbb{T}_n$, where $\mathbf{a} = (a_1, \ldots, a_n)^\top$ is a positive vector, $\mathbf{b} = (b_1, \ldots, b_m)^\top$ is a non-negative vector, and $\mathbf{s} = (s_1, \ldots, s_m)^\top$ with $0 := s_0 < 1 \leqslant s_1 < \cdots < s_m := n$.

According to Tian et al. (2003), if an n-dimensional random vector $\mathbf{X} \sim \mathrm{GD}_{n,m,\mathbf{s}}(\mathbf{a}, \mathbf{b})$ on $\mathbb{T}_n$, we can partition $\mathbf{X}$ into block vector $(\mathbf{X}_1, \ldots, \mathbf{X}_m)$ with the following stochastic representation:

$$\mathbf{X} = \begin{pmatrix} \mathbf{X}_1 \\ \vdots \\ \mathbf{X}_m \end{pmatrix} \overset{d}{=} \begin{pmatrix} R_1 \cdot \mathbf{Y}_1 \\ \vdots \\ R_m \cdot \mathbf{Y}_m \end{pmatrix}, \tag{2.10}$$

where (i) $\mathbf{R} = (R_1, \ldots, R_m)^\top \sim \mathrm{Dirichlet}(\beta_1, \ldots, \beta_m)$ with $\beta_j = \sum_{k=s_{j-1}+1}^{s_j} a_k + b_j$, $j = 1, \ldots, m$; (ii) $\mathbf{Y}_j \sim \mathrm{Dirichlet}(\mathbf{a}_j)$, $j = 1, \ldots, m$; and (iii) $\mathbf{R}$ and $(\mathbf{Y}_1, \ldots, \mathbf{Y}_m)$ are mutually independent. The notation $\mathbf{X} \overset{d}{=} \mathbf{Y}$ means that the random vectors $\mathbf{X}$ and $\mathbf{Y}$ have the same distribution.

Let $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_{16})^\top$, we have $\boldsymbol{\xi} | D_{obs} \sim \mathrm{GD}_{16,6,\mathbf{s}}(\mathbf{a}, \mathbf{b})$, where $\mathbf{a} = (n_1 + \alpha_{11}, n_2 + \alpha_{12}, n_3 + \alpha_{13}, n_4 + \alpha_{14}, \alpha_{21}, \alpha_{22}, \alpha_{23}, \alpha_{24}, \alpha_{31}, \alpha_{33}, \alpha_{32}, \alpha_{34}, \alpha_{41}, \alpha_{42}, \alpha_{43}, \alpha_{44})^\top$, $\mathbf{b} = (0, m_{12}, m_{34}, m_{13}, m_{24}, m_{1234})^\top$ and $\mathbf{s} = (4, 6, 8, 10, 12, 16)^\top$. Thus, we can generate i.i.d. samples from (2.10).

Next, from Table 3, note that θ_2, θ_3 and $\boldsymbol{\xi}$ have the following relationship:

$$\theta_2 = \sum_{i=1}^{4} \pi_{i2} = \xi_2 + \xi_6 + \xi_{11} + \xi_{14} \quad \text{and} \quad \theta_3 = \sum_{i=1}^{4} \pi_{i3} = \xi_3 + \xi_7 + \xi_{10} + \xi_{15}.$$

Thus, the Bayesian p-value can be computed from the following sampling-based method.

Step 1. Generate i.i.d. posterior sample $\boldsymbol{\xi}^{(1)}, \ldots, \boldsymbol{\xi}^{(G)}$ from $\mathrm{GD}_{16,6,\mathbf{s}}(\mathbf{a}, \mathbf{b})$, where $\boldsymbol{\xi}^{(g)} = (\xi_1^{(g)}, \ldots, \xi_{16}^{(g)})^\top$, $g = 1, \ldots, G$.
Step 2. Based on $\{\boldsymbol{\xi}^{(g)}\}_{g=1}^{G}$, calculate $\theta_2^{(g)} = \xi_2^{(g)} + \xi_6^{(g)} + \xi_{11}^{(g)} + \xi_{14}^{(g)}$ and $\theta_3^{(g)} = \xi_3^{(g)} + \xi_7^{(g)} + \xi_{10}^{(g)} + \xi_{15}^{(g)}$ for $g = 1, \ldots, G$.
Step 3. The Bayesian p-value is given by

$$p_{v1} = \frac{1}{G} \sum_{g=1}^{G} I\left(\theta_2^{(g)} \leqslant \theta_3^{(g)} - \delta_0 \right),$$

where $I(\cdot)$ is the indicator function.

2.3.2. The adjusted Bayesian p-value

Lindley (1957) pointed out a paradoxical result that even when a very small observed significance level is obtained, the posterior probability of being null can be close to one as long as the sample size is large enough. This observation motivated us to provide an alternative Bayesian test based on the method proposed in Yin (2012).

Specifically, let $\Delta = \theta_2 - \theta_3 - \delta_0$, the hypothesis in (2.1) becomes H_0: $\Delta \leqslant 0$ against H_1: $\Delta > 0$. Hence, a Bayesian measure for testing H_0 against H_1 can be defined by

$$p_{v2} = \Pr\left\{ \Delta - E(\Delta | D_{0bs}) \leqslant -|E(\Delta | D_{0bs})| \,\middle|\, D_{obs} \right\}, \tag{2.11}$$

where $E(\Delta | D_{obs})$ is the posterior expectation of Δ, which is given by

$$\begin{aligned} E(\Delta | D_{obs}) \quad &= E(\xi_2 | D_{obs}) + E(\xi_6 | D_{obs}) + E(\xi_{11} | D_{obs}) + E(\xi_{14} | D_{obs}) \\ - E(\xi_3 | D_{obs}) \; &- E(\xi_7 | D_{obs}) - E(\xi_{10} | D_{obs}) - E(\xi_{15} | D_{obs}) - \delta_0. \end{aligned} \tag{2.12}$$

Ng et al. (2008) showed that the expectation of any component of a random vector following a GDD has an explicit expression. From (6.5) in Ng et al. (2008), we obtain

$$\begin{cases} E(\xi_2|D_{obs}) = \dfrac{n_2+\alpha_{12}}{N+\alpha_{..}}, \\ E(\xi_6|D_{obs}) = \dfrac{\alpha_{22}(\alpha_{21}+\alpha_{22}+m_{12})}{(\alpha_{21}+\alpha_{22})(N+\alpha_{..})}, \\ E(\xi_{11}|D_{obs}) = \dfrac{\alpha_{32}(\alpha_{32}+\alpha_{34}+m_{24})}{(\alpha_{32}+\alpha_{34})(N+\alpha_{..})}, \\ E(\xi_{14}|D_{obs}) = \dfrac{\alpha_{42}(\alpha_{4.}+m_{1234})}{\alpha_{4.}(N+\alpha_{..})}, \\ E(\xi_3|D_{obs}) = \dfrac{n_3+\alpha_{13}}{N+\alpha_{..}}, \\ E(\xi_7|D_{obs}) = \dfrac{\alpha_{23}(\alpha_{23}+\alpha_{24}+m_{24})}{(\alpha_{23}+\alpha_{24})(N+\alpha_{..})}, \\ E(\xi_{10}|D_{obs}) = \dfrac{\alpha_{33}(\alpha_{31}+\alpha_{33}+m_{13})}{(\alpha_{31}+\alpha_{33})(N+\alpha_{..})}, \\ E(\xi_{15}|D_{obs}) = \dfrac{\alpha_{43}(\alpha_{4.}+m_{1234})}{\alpha_{4.}(N+\alpha_{..})}, \end{cases} \tag{2.13}$$

where $N = \sum_{i=1}^{4} n_i + m_{12} + m_{34} + m_{13} + m_{24} + m_{1234}$, $\alpha_{4.} = \sum_{j=1}^{4}\alpha_{4j}$ and $\alpha_{..} = \sum_{i=1}^{4}\sum_{j=1}^{4}\alpha_{ij}$. In other words, we can compute $E(\Delta|D_{obs})$ exactly.

We propose the following sampling-based method for calculating the posterior probability p_{v2}:

Step 1. Generate i.i.d. samples $\boldsymbol{\xi}^{(1)}, \ldots, \boldsymbol{\xi}^{(G)}$ from $\mathrm{GD}_{16,6,\mathbf{s}}(\mathbf{a}, \mathbf{b})$, where $\boldsymbol{\xi}^{(g)} = (\xi_1^{(g)}, \ldots, \xi_{16}^{(g)})^\top$, $g = 1, \ldots, G$.
Step 2. Based on $\{\boldsymbol{\xi}^{(g)}\}_{g=1}^{G}$, calculate $\theta_2^{(g)} = \xi_2^{(g)} + \xi_6^{(g)} + \xi_{11}^{(g)} + \xi_{14}^{(g)}$ and $\theta_3^{(g)} = \xi_3^{(g)} + \xi_7^{(g)} + \xi_{10}^{(g)} + \xi_{15}^{(g)}$ for $g = 1, \ldots, G$.
Step 3. Compute the posterior expectation $E(\Delta|D_{obs})$ through (2.12) and (2.13).
Step 4. The Bayesian p-value in (2.11) can be obtained through

$$p_{v2} = \frac{1}{G}\sum_{g=1}^{G} I\left(\theta_2^{(g)} - \theta_3^{(g)} - \delta_0 - E(\Delta|D_{obs}) \leqslant -|E(\Delta|D_{obs})|\right).$$

3. Bayesian credible intervals for risk difference

The construction of confidence intervals for parameters of interest in matched-pair design has been extensively explored in the literature (Liu et al., 2002; Newcombe, 2003). In the Bayesian framework, for two binary outcomes in a 2×2 contingency table, Agresti and Min (2005) and Hashemi et al. (1998) examined Bayesian credible intervals for parameters such as proportion difference, relative risk and odds ratio. For matched-pair designs with incomplete observations under MAR, Shi and Bai (2008, 2009) explored Bayesian credible intervals. For incomplete paired-data under non-ignorable missing mechanism, Bayesian credible intervals are less explored in the literature. In this section, we shall consider the construction of two Bayesian credible intervals (i.e., equal-tail credible interval and HPD interval) for risk difference in matched-pair designs with incomplete observations under non-ignorable missing mechanism.

In Section 2.3.1, we have generated i.i.d. posterior samples $\{\theta_2^{(g)}, \theta_3^{(g)}\}_{g=1}^{G}$, which can be used to calculate the risk difference $\delta = \theta_2 - \theta_3$. Consequently, the $(1-\alpha)100\%$ equal-tail credible interval for δ is given by

$$[\hat{\delta}_{\mathrm{L}},\ \hat{\delta}_{\mathrm{U}}], \tag{3.1}$$

where $\hat{\delta}_{\mathrm{L}}$ and $\hat{\delta}_{\mathrm{U}}$ are the $100(\alpha/2)$ and $100(1-\alpha/2)$ percentiles of $\{\delta^{(g)}\}_{g=1}^{G}$, where $\delta^{(g)} = \theta_2^{(g)} - \theta_3^{(g)}$.

To find the $(1-\alpha)100\%$ highest posterior density (HPD) interval for δ, we first sort the i.i.d. posterior samples $\{\delta^{(g)}\}_{g=1}^{G}$ to obtain their order statistics, denoted by $\{\delta_{(g)}\}_{g=1}^{G}$. The $(1-\alpha)100\%$ HPD interval for δ can be obtained by minimizing the width of the following $(1-\alpha)100\%$ Bayesian credible intervals of δ:

$$\left[\delta_{(g)},\ \delta_{(\lceil G(1-\alpha)\rceil+g)}\right], \quad g = 1, \ldots, \lceil G\alpha\rceil, \tag{3.2}$$

where $\lceil r\rceil$ is the largest integer not greater than r.

To demonstrate, let $G = 1000$ and $\alpha = 0.05$. We can then have 50 Bayesian credible intervals for δ with 95% coverage, i.e., $[\delta_{(g)},\ \delta_{(950+g)}]$, $g = 1, \ldots, 50$. The 95% HPD interval for δ is $[\delta_{(g^*)},\ \delta_{(950+g^*)}]$, whose width $\delta_{(950+g^*)} - \delta_{(g^*)} = \min_{1\leqslant g\leqslant 50}\{\delta_{(950+g)} - \delta_{(g)}\}$ is the shortest among the 50 Bayesian credible intervals. In fact, the 95% equal-tail credible interval for δ is given by $[\delta_{(25)}, \delta_{(975)}]$.

Table 4
Empirical type I error based on 10,000 replications and 5000 sampling-based samples.

N	π	Results for different priors:					
		$\alpha = 0.5 \cdot \mathbb{1}_{16}$		$\alpha = 1.0 \cdot \mathbb{1}_{16}$		$\alpha = 1.5 \cdot \mathbb{1}_{16}$	
		p_{v1}	p_{v2}	p_{v1}	p_{v2}	p_{v1}	p_{v2}
30	Case a	0.034	0.109	0.018	0.071	0.025	0.051
	Case b	0.024	0.057	0.018	0.047	0.018	0.040
	Case c	0.011	0.043	0.014	0.038	0.008	0.028
	Case d	0.012	0.040	0.01	0.034	0.011	0.029
50	Case a	0.028	0.105	0.028	0.103	0.025	0.077
	Case b	0.023	0.062	0.024	0.060	0.011	0.046
	Case c	0.007	0.036	0.018	0.042	0.01	0.034
	Case d	0.01	0.035	0.009	0.037	0.007	0.027
100	Case a	0.039	0.137	0.028	0.136	0.027	0.122
	Case b	0.025	0.053	0.016	0.048	0.023	0.053
	Case c	0.005	0.012	0.008	0.035	0.01	0.028
	Case d	0.009	0.032	0.011	0.026	0.007	0.038

4. Simulation studies

In this section, we conduct simulation studies to investigate the finite sample performance of the proposed method. Let $\mathbf{J}_4$ be a 4×4 matrix where every element is equal to 1. The following prior distributions are considered: (i) $\boldsymbol{\alpha} = 0.5 \times \mathbf{J}_4$ corresponds to Jeffreys non-informative prior. With complete data, this produces a posterior distribution, where the posterior mean minimizes the Bayes risk under the quadratic loss. (ii) $\boldsymbol{\alpha} = \mathbf{J}_4$ gives a posterior mode which is the usual MLE with no missing data. (iii) $\boldsymbol{\alpha} = 1.5 \times \mathbf{J}_4$ gives a posterior mode which is the usual MLE after adding 0.5 to each cell count in the 2×2 table with no missing data.

To eliminate nuisance parameters, the p-value we computed is based on the posterior distribution of nuisance parameters. This idea showed up in Guttman (1967) and Rubin (1984). There is an apparent "double use" of data, first to convert prior into a proper posterior distribution, and then to compute the tail area based on observed value of the test statistic. To circumvent this "double usage" of data, Bayarri and Berger (2000) proposed conditional predictive p value and the partial posterior predictive p value, which follow $\mathcal{U}(0, 1)$ under certain assumptions. It is too complicated for us to use the conditional predictive p value and the partial posterior predictive p value in our set up. To evaluate how close our p value is with the $\mathcal{U}(0, 1)$, in Fig. 1 and Fig. 2, we plot their cumulative distribution functions for four cases (a,b,c,d) and the prior $\alpha = 0.5\mathbf{1}_{16}$ used in the simulation studies empirically and put them in Appendix A. It turns out that the cdf of p_{v1} is fairly close to $\mathcal{U}(0, 1)$ and p_{v2} can be anti-conservative most of the time.

To compute the type I error rate, we set up the cell probabilities to satisfy $\theta_2 = \theta_3 - 0.01$ (i.e., $\sum_{i=1}^{4}\pi_{i2} = \sum_{i=1}^{4}\pi_{i3} - 0.01$) and $\sum_{i=1}^{4}\sum_{j=1}^{4}\pi_{ij} = 1$ as in Table B.1 in the Appendix. We consider three scenarios: no missing data (case a), MAR (case b) and MNAR (case c and d). We consider sample size $N = 30, 50, 100$.

Given N and cell probabilities $\boldsymbol{\pi}$, we generate

$$(n_1, n_2, n_3, n_4, m_{12}, m_{34}, m_{13}, m_{24}, m_{1234}) \sim \text{Multinomial}(N; \pi_{11}, \ldots, \pi_{14},$$
$$\pi_{21} + \pi_{22}, \pi_{23} + \pi_{24}, \pi_{31} + \pi_{33}, \pi_{32} + \pi_{34}, \pi_{41} + \cdots + \pi_{44})$$

and obtain the observed data $D_{obs} = \{n_1, \ldots, n_4; m_{12}, m_{34}, m_{13}, m_{24}, m_{1234}\}$. Based on D_{obs} and $\boldsymbol{\alpha}$, we generate $G = 5000$ i.i.d. posterior samples $\{\boldsymbol{\xi}^{(g)}\}_{g=1}^{G}$, and calculate two Bayesian p-values p_{v1} and p_{v2}. Repeat this process 10,000 times, we get the empirical type I error rate by counting the percentage of p-values that are smaller than 0.05. Bayarri and Berger (2000) showed that p-values obtained from the posterior distribution after plugging in corresponding estimates are conservative. The results are summarized in Table 4. The nominal type I error control is 0.05.

From Table 4, we observe that the method based on p_{v1} controls type I error rate conservatively over all scenarios. Method based on p_{v2} has type I error rate close to the nominal 0.05 when there are missing data and has inflated type I error rate when there is no missing data.

We next investigate power. We set up the cell probabilities to satisfy $\theta_2 > \theta_3 - 0.01$ (i.e., $\sum_{i=1}^{4}\pi_{i2} > \sum_{i=1}^{4}\pi_{i3} - 0.01$, and $\sum_{i=1}^{4}\sum_{j=1}^{4}\pi_{ij} = 1$) as in Table B.2 in the Appendices A and B. We consider three scenarios: no missing data (case a1), MAR (case b1) and MNAR (case c1 and d1). Sample size $N = 30, 50, 100$. The rest are exactly the same as in the calculation of type I error rate. The results are summarized in Table 5.

From Table 5, we observe that the two approaches have very similar power across all priors, sample sizes and missingness mechanism. Jeffreys non-informative prior $\boldsymbol{\alpha} = 0.5 \times \mathbf{J}_4$ produces higher power than other priors for both approaches. The proposed methods have higher power under MNAR than under MAR. Power improves with increased sample size.

To investigate the performance of the proposed Bayesian credible intervals for the risk difference, we compute their empirical coverage probabilities (ECPs) and empirical confidence widths (ECWs) through simulations. Table B.3 in the Appendices A and B lists the configuration of cell probabilities $\boldsymbol{\pi}$, which satisfies $\delta = -0.25, 0, 0.25$ and $\sum_{i=1}^{4}\sum_{j=1}^{4}\pi_{ij} = 1$.

Table 5
Empirical power based on 10,000 replications and 5000 sampling samples.

N	$\boldsymbol{\pi}$	Results for the different priors:					
		$\boldsymbol{\alpha} = 0.5 \cdot \mathbb{1}_{16}$		$\boldsymbol{\alpha} = 1.0 \cdot \mathbb{1}_{16}$		$\boldsymbol{\alpha} = 1.5 \cdot \mathbb{1}_{16}$	
		p_{v1}	p_{v2}	p_{v1}	p_{v2}	p_{v1}	p_{v2}
30	Case a_1	0.49	0.49	0.444	0.453	0.380	0.382
	Case b_1	0.111	0.112	0.094	0.096	0.070	0.071
	Case c_1	0.417	0.418	0.411	0.411	0.395	0.398
	Case d_1	0.989	0.989	0.975	0.976	0.973	0.975
50	Case a_1	0.675	0.676	0.640	0.645	0.632	0.633
	Case b_1	0.144	0.145	0.170	0.173	0.175	0.175
	Case c_1	0.665	0.665	0.670	0.672	0.675	0.676
	Case d_1	0.997	0.997	0.999	0.999	0.999	0.999
100	Case a_1	0.923	0.923	0.899	0.901	0.916	0.920
	Case b_1	0.246	0.250	0.277	0.278	0.307	0.307
	Case c_1	0.898	0.899	0.911	0.911	0.926	0.927
	Case d_1	0.997	0.998	0.999	0.999	0.999	0.999

We consider three scenarios: no missing data (case a2), MAR (case b2) and MNAR (case c2). Sample size $N = 30, 50, 100, 400$. Given N and cell probabilities $\boldsymbol{\pi}$, data generation follows the same as in the type I error simulation.

We generate $G = 5000$ i.i.d. posterior sample $\{\boldsymbol{\xi}^{(g)}\}_{g=1}^{G}$, and calculate

$$\begin{aligned}\delta^{(g)} &= \theta_2^{(g)} - \theta_3^{(g)} \\ &= \xi_2^{(g)} + \xi_6^{(g)} + \xi_{11}^{(g)} + \xi_{14}^{(g)} - \xi_3^{(g)} - \xi_7^{(g)} - \xi_{10}^{(g)} - \xi_{15}^{(g)}.\end{aligned}$$

Denote the order statistic as $\{\delta_{(g)}\}_{g=1}^{G}$. At significance level 0.05, we can have 250 Bayesian credible intervals for δ with 95% coverage, i.e., $[\delta_{(g)}, \delta_{(4750+g)}]$, $g = 1, \ldots, 250$. The 95% equal-tail credible interval for δ is $[\delta_{(125)}, \delta_{(4875)}]$ and the 95% HPD interval for δ is $[\delta_{(g^*)}, \delta_{(4750+g^*)}]$, whose width $\delta_{(4750+g^*)} - \delta_{(g^*)} = \min_{1 \leqslant g \leqslant 250}\{\delta_{(4750+g)} - \delta_{(g)}\}$ is the shortest among the 250 Bayesian credible intervals.

Let $M = 10{,}000$ be the number of repetitions used in the simulation and $[\delta_{\mathrm{L}}^{(m)}, \ \delta_{\mathrm{U}}^{(m)}]$ be the 95% Bayesian credible interval of δ obtained from the mth simulation run. The empirical coverage probability is calculated through

$$\mathrm{ECP} = \frac{1}{M}\sum_{m=1}^{M} I\{\delta \in [\delta_{\mathrm{L}}^{(m)}, \ \delta_{\mathrm{U}}^{(m)}]\},$$

where $\delta = \sum_{i=1}^{4}(\pi_{i2} - \pi_{i3})$ is completely determined by $\boldsymbol{\pi}$. The empirical confidence width is obtained through

$$\mathrm{ECW} = \frac{1}{M}\sum_{m=1}^{M}[\delta_{\mathrm{U}}^{(m)} - \delta_{\mathrm{L}}^{(m)}].$$

Table 6 summarizes simulation results of empirical coverage probabilities. From Table 6, we observe that posterior credible interval and HPD have similar performance across different priors. Empirical coverage probabilities are close to the nominal 0.95 for complete data. With missing data, coverage probabilities are greater than 95% regardless of MAR or MNAR.

Table 7 summarizes simulation results of empirical confidence widths. We observe that empirical confidence widths based on posterior credible interval and HPD are similar. Prior $\boldsymbol{\alpha} = 1.5 \times \mathbf{J}_4$ produces the narrowest empirical confidence width. Empirical confidence widths for complete data is the shortest. As sample size increases, the empirical confidence width gets smaller.

5. Two real examples

5.1. A study of medical malpractice

We revisit the study of medical malpractice data example introduced in Section 1. From Table 1, we have $n_1 = 26, n_2 = 1, n_3 = 5, n_4 = 18, m_{12} = 2, m_{34} = 9, m_{13} = 4, m_{24} = 4, m_{1234} = 0$. We use the proposed Bayesian p-values to test hypothesis (2.1). We also analyze this dataset under the MAR assumption. The p-values are reported in Table 8 and Bayesian posterior intervals are reported in Table 9. We can see that under MAR, when prior $\boldsymbol{\alpha} = 0.5 \times \mathbf{J}_4$, the p-values are smaller than 0.05 and the Bayesian posterior intervals do not contain 0, suggesting strong evidence to reject the null. Still under MAR, if prior $\boldsymbol{\alpha} = 1.0 \times \mathbf{J}_4$ or $1.5 \times \mathbf{J}_4$, the p-values are borderline significant and Bayesian posterior intervals contain 0. These suggest weak evidence to reject the null. On the other hand, under MNAR, p-values are greater than 0.05 and the Bayesian posterior intervals all contain 0, consistently showing the we do not reject the null.

Table 6
Empirical coverage probabilities based on 10,000 replications and 5000 sampling samples.

N	π	Results for the different priors:					
		$\alpha = 0.5 \cdot \mathbb{1}_{16}$		$\alpha = 1.0 \cdot \mathbb{1}_{16}$		$\alpha = 1.5 \cdot \mathbb{1}_{16}$	
		HPD	Credible interval	HPD	Credible interval	HPD	Credible interval
30	Case a_2	0.953	0.962	0.980	0.982	0.987	0.987
	Case b_2	0.992	0.992	0.996	0.995	0.956	0.956
	Case c_2	0.998	0.998	0.990	0.991	0.951	0.952
50	Case a_2	0.947	0.954	0.972	0.974	0.981	0.982
	Case b_2	0.998	0.997	0.992	0.992	0.978	0.977
	Case c_2	1.000	1.000	0.993	0.993	0.982	0.981
100	Case a_2	0.946	0.951	0.950	0.953	0.967	0.965
	Case b_2	0.999	0.999	1.000	1.000	0.997	0.996
	Case c_2	1.000	1.000	0.998	0.998	0.996	0.996
400	Case a_2	0.952	0.954	0.952	0.951	0.962	0.963
	Case b_2	0.999	0.999	1.000	1.000	0.997	0.996
	Case c_2	1.000	1.000	1.000	1.000	1.000	1.000

Table 7
Empirical confidence widths based on 10,000 replications and 5000 sampling samples.

N	π	Results for the different priors:					
		$\alpha = 0.5 \cdot \mathbb{1}_{16}$		$\alpha = 1.0 \cdot \mathbb{1}_{16}$		$\alpha = 1.5 \cdot \mathbb{1}_{16}$	
		HPD	Credible interval	HPD	Credible interval	HPD	Credible interval
30	Case a_2	0.437	0.438	0.400	0.401	0.371	0.373
	Case b_2	0.510	0.512	0.445	0.446	0.402	0.403
	Case c_2	0.513	0.515	0.446	0.448	0.402	0.404
50	Case a_2	0.357	0.359	0.334	0.336	0.318	0.319
	Case b_2	0.449	0.451	0.393	0.395	0.358	0.360
	Case c_2	0.449	0.450	0.392	0.394	0.358	0.359
100	Case a_2	0.263	0.264	0.254	0.255	0.246	0.247
	Case b_2	0.386	0.388	0.333	0.335	0.304	0.305
	Case c_2	0.386	0.387	0.334	0.335	0.305	0.306
400	Case a_2	0.137	0.137	0.135	0.136	0.134	0.134
	Case b_2	0.385	0.388	0.333	0.335	0.304	0.305
	Case c_2	0.323	0.324	0.270	0.268	0.236	0.237

Table 8
Results for testing (2.1) in the study of medical malpractice example.

Prior	$\alpha = 0.5 \cdot \mathbb{1}_{16}$		$\alpha = 1.0 \cdot \mathbb{1}_{16}$		$\alpha = 1.5 \cdot \mathbb{1}_{16}$	
	p_{v1}	p_{v2}	p_{v1}	p_{v2}	p_{v1}	p_{v2}
p-value under NMAR	0.899	0.109	0.899	0.106	0.892	0.111
p-value under MAR	0.024	0.038	0.032	0.046	0.04	0.053

Table 9
95% Bayesian intervals for δ in the study of medical malpractice example.

Missing mechanism	Prior	Type of Bayesian intervals					
		HPD			Credible interval		
		Lower	Upper	Width	Lower	Upper	Width
NMAR	$\alpha = 0.5 \cdot \mathbb{1}_{16}$	−0.249	0.0507	0.299	−0.252	0.048	0.299
	$\alpha = 1.0 \cdot \mathbb{1}_{16}$	−0.228	0.045	0.273	−0.227	0.046	0.273
	$\alpha = 1.5 \cdot \mathbb{1}_{16}$	−0.212	0.046	0.258	−0.211	0.047	0.258
MAR	$\alpha = 0.5 \cdot \mathbb{1}_{16}$	−0.218	−0.0320	0.186	−0.218	−0.032	0.186
	$\alpha = 1.0 \cdot \mathbb{1}_{16}$	−0.172	0.0133	0.186	−0.170	0.013	0.182
	$\alpha = 1.5 \cdot \mathbb{1}_{16}$	−0.160	0.014	0.174	−0.158	0.013	0.171

5.2. A crime survey example

We illustrate the proposed method with data from a crime survey study, which was reported in Kadane (1985) and Tian et al. (2003). We have $n_1 = 392$, $n_2 = 55$, $n_3 = 76$, $n_4 = 38$, $m_{12} = 33$, $m_{34} = 9$, $m_{13} = 31$, $m_{24} = 7$, $m_{1234} = 115$. The p-values are summarized in Table 10 and Bayesian posterior intervals are summarized in Table 11. Under MAR, p-values show evidence to reject the null yet the Bayesian posterior intervals all contain 0 regardless of the prior. Under MNAR,

Table 10
Results for testing (2.1) in the study of crime survey.

Prior	$\boldsymbol{\alpha} = 0.5 \cdot \mathbb{1}_{16}$		$\boldsymbol{\alpha} = 1.0 \cdot \mathbb{1}_{16}$		$\boldsymbol{\alpha} = 1.5 \cdot \mathbb{1}_{16}$	
	p_{v1}	p_{v2}	p_{v1}	p_{v2}	p_{v1}	p_{v2}
p-value under NMAR	0.664	0.664	0.694	0.694	0.716	0.716
p-value under MAR	0.030	0.031	0.037	0.041	0.040	0.046

Table 11
95% Bayesian intervals for δ in the study of crime survey.

Missing mechanism	Prior	Type of Bayesian intervals					
		HPD			Credible interval		
		Lower	Upper	Width	Lower	Upper	Width
NMAR	$\boldsymbol{\alpha} = 0.5 \cdot \mathbb{1}_{16}$	−0.163	0.109	0.272	−0.163	0.109	0.272
	$\boldsymbol{\alpha} = 1.0 \cdot \mathbb{1}_{16}$	−0.136	0.082	0.217	−0.137	0.080	0.217
	$\boldsymbol{\alpha} = 1.5 \cdot \mathbb{1}_{16}$	−0.121	0.067	0.188	−0.121	0.067	0.188
MAR	$\boldsymbol{\alpha} = 0.5 \cdot \mathbb{1}_{16}$	−0.102	0.003	0.105	−0.102	0.003	0.104
	$\boldsymbol{\alpha} = 1.0 \cdot \mathbb{1}_{16}$	−0.082	0.003	0.085	−0.082	0.003	0.085
	$\boldsymbol{\alpha} = 1.5 \cdot \mathbb{1}_{16}$	−0.076	0.002	0.078	−0.076	0.002	0.078

Table B.1
Cell probabilities for calculating type I error.

Case a:	$\boldsymbol{\pi} = \left(\frac{1}{4}, \frac{6}{25}, \frac{13}{50}, \frac{1}{4}, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0\right)^\top$
Case b:	$\boldsymbol{\pi} = \left(\frac{1}{8}, \frac{1}{4}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}, 0, 0, \frac{1}{16}, 0, \frac{21}{400}, \frac{29}{400}, 0, 0, 0, 0, 0\right)^\top$
Case c:	$\boldsymbol{\pi} = \left(\frac{1}{32}, \frac{1}{4}, \frac{1}{4}, \frac{1}{32}, \frac{1}{32}, \frac{1}{32}, \frac{1}{32}, \frac{1}{32}, \frac{1}{32}, \frac{21}{400}, \frac{29}{400}, \frac{1}{32}, \frac{1}{32}, \frac{1}{32}, \frac{1}{32}, \frac{1}{32}\right)^\top$
Case d:	$\boldsymbol{\pi} = \left(\frac{1}{16}, \frac{1}{4}, \frac{1}{4}, \frac{1}{16}, \frac{1}{32}, \frac{1}{32}, \frac{1}{32}, \frac{1}{32}, \frac{1}{32}, \frac{17}{800}, \frac{33}{800}, \frac{1}{32}, \frac{1}{32}, \frac{1}{32}, \frac{1}{32}, \frac{1}{32}\right)^\top$

p-values are greater than 0.05 and the Bayesian posterior intervals all contain 0 regardless of the prior, providing consistent evidence not to reject the null. We noted that the interval width is shorter under MAR than MNAR, which is expected.

6. Concluding remarks

In this paper, we study non-inferiority testing with paired data under non-ignorable missing mechanism. We present two Bayesian p-value approaches and provide Bayesian credible intervals and highest posterior density intervals for the risk difference. Numerical studies show that the proposed methods can control type I error rate and have decent power. Empirical coverage probabilities always exceed the nominal 95%. We find Jeffreys non-informative prior consistently has good performance and recommend its usage in practice.

Acknowledgments

The authors would like to thank the editor, the associate editor and two referees for insightful comments that help to streamline the paper. The research of Hui-qiong Li is partially supported by the Natural Science Foundation of China (11561075) and grants from the Science Foundation of Yunnan Province (2016FB005). Guo-liang Tian is partly supported by the Natural Science Foundation of China (11771199).

Appendix A. Cumulative distribution function of p_{v1} and p_{v2}

See Figs. 1 and 2.

Appendix B. This appendix contains cell probabilities to generate data in the simulation studies

See Tables B.1–B.3.

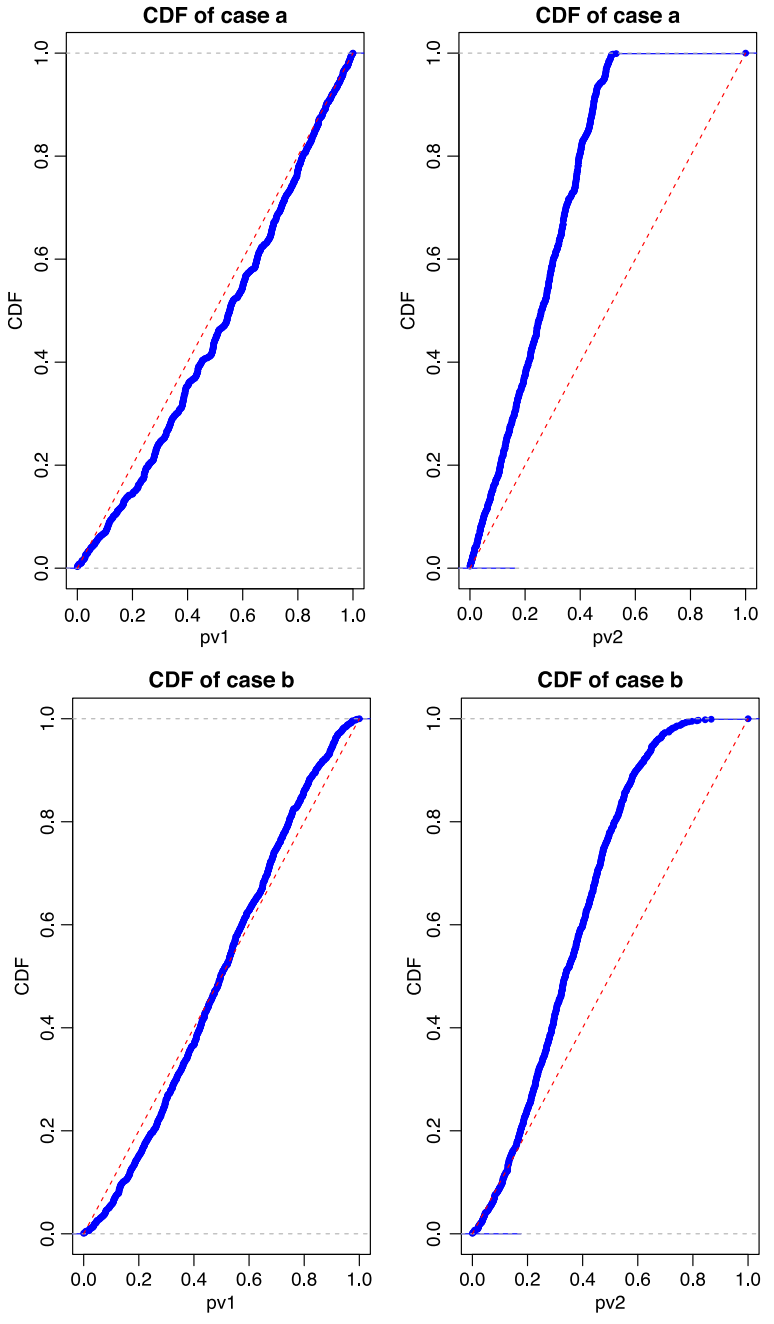

Fig. 1. cdf under the prior $\alpha = 0.5\mathbf{1}_{16}$.

Table B.2
Cell probabilities for calculating the power.

Case a_1:	$\boldsymbol{\pi} = \left(\frac{1}{8}, \frac{1}{2}, \frac{1}{4}, \frac{1}{8}, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0\right)^\top$
Case b_1:	$\boldsymbol{\pi} = \left(\frac{1}{6}, \frac{1}{4}, \frac{1}{12}, \frac{1}{12}, 0, \frac{1}{12}, \frac{1}{12}, 0, \frac{1}{12}, 0, \frac{1}{12}, \frac{1}{12}, 0, 0, 0, 0\right)^\top$
Case c_1:	$\boldsymbol{\pi} = \left(\frac{1}{32}, \frac{3}{8}, \frac{1}{8}, \frac{1}{32}, \frac{1}{32}, \frac{1}{16}, 0, \frac{1}{32}, \frac{1}{32}, \frac{1}{16}, \frac{1}{16}, \frac{1}{32}, \frac{1}{32}, \frac{1}{16}, 0, \frac{1}{32}\right)^\top$
Case d_1:	$\boldsymbol{\pi} = \left(\frac{1}{16}, \frac{1}{2}, 0, \frac{1}{16}, 0, \frac{1}{16}, \frac{1}{32}, \frac{1}{32}, 0, \frac{1}{16}, \frac{1}{16}, \frac{1}{32}, \frac{1}{32}, \frac{1}{16}, 0, \frac{1}{32}\right)^\top$

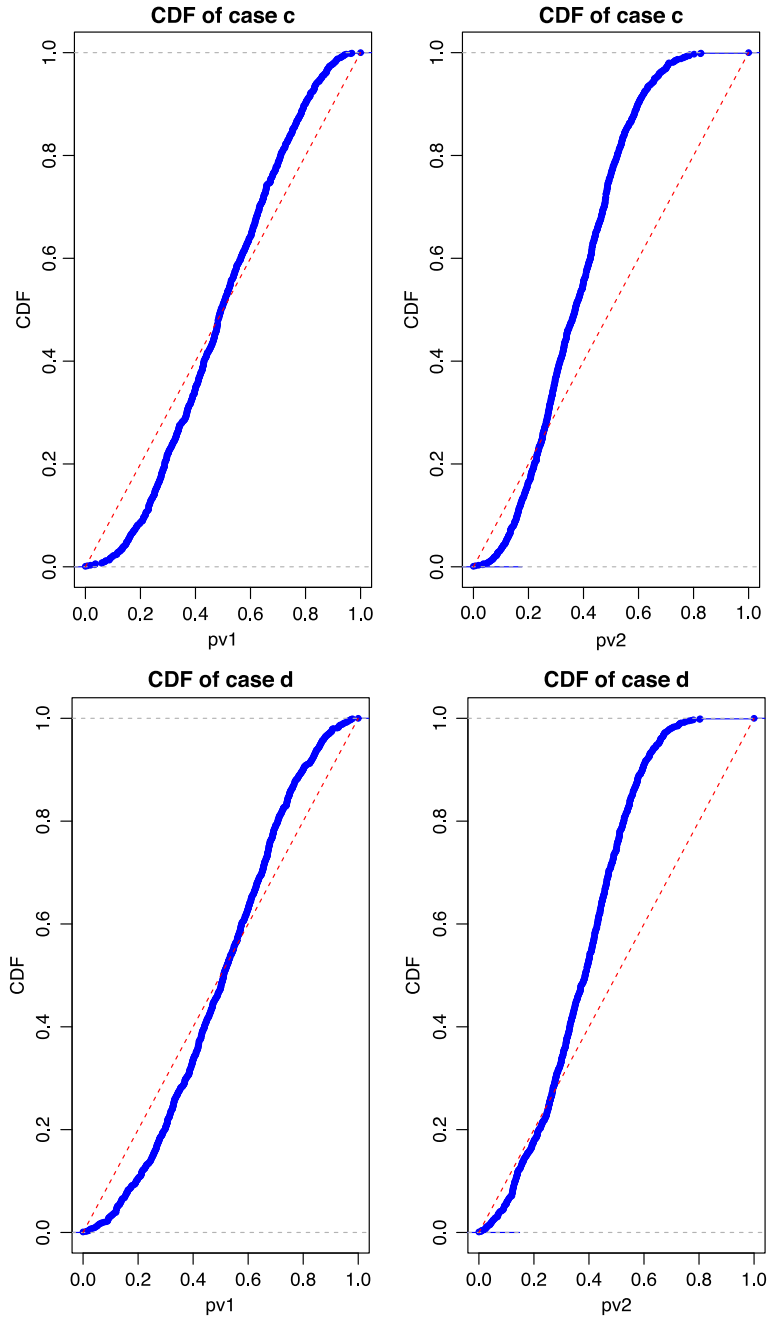

Fig. 2. cdf under the prior $\alpha = 0.5\mathbf{1}_{16}$.

Table B.3
Cell probabilities for empirical coverage probabilities and empirical confidence widths.

Case a_2:	$\delta = 0$	$\boldsymbol{\pi} = \left(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0\right)^\top$
Case b_2:	$\delta = -0.25$	$\boldsymbol{\pi} = \left(\frac{1}{8}, 0, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}, 0, 0, \frac{1}{16}, \frac{1}{4}, \frac{1}{16}, \frac{1}{16}, 0, 0, 0, 0, 0\right)^\top$
Case c_2:	$\delta = 0.25$	$\boldsymbol{\pi} = \left(\frac{1}{32}, \frac{1}{4}, 0, \frac{9}{32}, \frac{1}{32}, \frac{1}{32}, \frac{1}{32}, \frac{1}{32}, \frac{1}{32}, \frac{1}{16}, \frac{1}{16}, \frac{1}{32}, \frac{1}{32}, \frac{1}{32}, \frac{1}{32}, \frac{1}{32}\right)^\top$

References

Agresti, A., Min, Y.Y., 2005. Frequentist performance of Bayesian confidence intervals for comparing proportions in 2 × 2 tables. Biometrics 61, 515–523.
Altham, P.M.E., 1969. Exact Bayesian analysis of a 2 × 2 contingency table, and Fisher's exact significance test. J. R. Stat. Soc. Ser. B 31, 261–269.
Altham, P.M.E., 1971. The analysis of matched proportions. Biometrika 58, 561–576.
Altham, P.M.E., Hankin, R.K.S., 2010. Correspondence: Using recently developed software on a 2 × 2 table of matched pairs with incompletely classified data. J. Roy. Statist. Soc. Ser. C 59, 377–379.
Bayarri, M.J., Berger, J.O., 2000. *P* values for composite null models. J. Amer. Statist. Assoc. 95, 1127–1142.
Chan, I.S.F., Tang, N.S., Tang, M.L., Chan, P.S., 2003. Statistical analysis of non-inferiority trials with a rate ratio in small-sample matched-pair designs. Biometrics 59, 1170–1177.
Choi, S.C., Stablein, D.M., 1982. Practical tests for comparing two proportions with incomplete data. Stat. Med. 7, 929–939.
Choi, S.C., Stablein, D.M., 1988. Comparing incomplete paired binomial data under non-random mechanisms. Stat. Med. 7, 929–939.
Ekbohm, G., 1982. On testing the equality of proportions in the paired case with incomplete data. Psychometrika 47, 115–118.
Greenberg, C.C., Regenbogen, S.E., Studdert, D.M., Lipsitz, S.R., Rogers, S.O., Zinner, M.J., Gawande, A.A., 2007. Patterns of communication breakdowns resulting in injury to surgical patients. J. Amer. College Surg. 204, 533–540.
Guttman, I., 1967. The use of the concept of a future observation in goodness-of-fit problems. J. Roy. Statist. Soc. Ser. B 29, 83–100.
Hashemi, L., Nandram, B., Goldberg, R., 1998. Bayesian analysis for a single 2 × 2 contingency table. Stat. Med. 16, 1311–1328.
Kadane, J.B., 1985. Is victimization chronic? a Bayesian analysis of a legal procedure. J. Amer. Statist. Assoc. 78, 544–552.
Kass, R.E., Raftery, A.E., 1995. Bayes factors. J. Amer. Statist. Assoc. 90, 773–795.
Konietschke, F., Harrar, S.W., Lange, K., Brunner, E., 2012. Ranking procedures for matched pairs with missing data—asymptotic theory and a small sample approximation. Comput. Statist. Data Anal. 56, 1090–1102.
Lin, Y., Lipsitz, S., Sinha, D., Gawande, A.A., Regenbogen, S.E., Greenberg, C.C., 2009. Using Bayesian *p*-values in a 2×2 table of matched pairs with incompletely classified data. J. Roy. Statist. Soc. Ser. C 58, 237–246.
Lindley, D.V., 1957. A statistical paradox. Biometrika 44, 187–192.
Little, R.J.A., Rubin, D.B., 2002. Statistical Analysis with Missing Data, second ed.. Wiley, New York.
Liu, J.P., Hsueh, H.M., Hsieh, E., Chen, J.J., 2002. Tests for equivalence or non-inferiority for paired binary data. Stat. Med. 21, 231–245.
Mosteller, F., 1952. Some statistical problems in measuring the subjective response to drugs. Biometrics 8, 220–226.
Nandram, B., Choi, J.W., 2002. A Bayesian analysis of a proportion under non-ignorable non-response. Stat. Med. 21, 1189–1212.
Newcombe, R.G., 2003. Confidence intervals for the mean of a variable taking the values 0, 1 and 2. Stat. Med. 22, 2085–2086.
Ng, K.W., Tang, M.L., Tan, M., Tian, G.L., 2008. Grouped Dirichlet distribution: A new tool for incomplete categorical data analysis. J. Multivariate Anal. 99 (3), 490–509.
Rubin, D.B., 1984. Bayesian justifiable and relevant frequency calculations for the applied statistician. Ann. Statist. 12, 1151–1172.
Shi, L., Bai, P., 2008. Bayesian confidence interval for the difference of two proportions in the matched-paired design. Comm. Statist. Theory Methods 37, 2034–2051.
Shi, L., Bai, P., 2009. Bayesian confidence interval for the ratio of marginal probabilities in the matched-paired design. Comm. Statist. Theory Methods 38, 1300–1316.
Tang, M.L., Li, H.Q., Chan, I.S.F., Tian, G.L., 2011. On confidence interval construction for establishing equivalence of two binary-outcome treatments in matched-pair studies in the presence of incomplete data. Statist. Biosci. 3, 223–249.
Tang, M.L., Ling, M.H., Tian, G.L., 2009. Exact and approximate unconditional confidence intervals for proportion difference in the presence of incomplete data. Stat. Med. 28, 625–641.
Tang, M.L., Tang, N.S., 2004. Exact tests for comparing two paired proportions with incomplete data. Biom. J. 46, 72–82.
Tang, N.S., Tang, M.L., Chan, I.S.F., 2003. On tests of equivalence via non-unity relative risk for matched-pair design. Stat. Med. 22, 1217–1233.
Tango, T., 1998. Equivalence test and confidence interval for the difference in proportions for the paired-sample design. Stat. Med. 17, 891–908.
Tian, G.L., Ng, K.W., Geng, Z., 2003. Bayesian computation for contingency tables with incomplete cell-counts. Statist. Sinica 13, 189–206.
Yin, Y.L., 2012. A new Bayesian procedure for testing point null hypotheses. Comput. Statist. 27, 237–249.