9-7-2006

Multiple imputation for the comparison of two screening tests in two-phase Alzheimer studies

Ofer Harel
University of Connecticut, oharel@stat.uconn.edu

Xiao-Hua Zhou
University of Washington, azhou@u.washington.edu

STATISTICS IN MEDICINE
Statist. Med. 2000; **00**:1–6 *Prepared using* simauth.cls *[Version: 2002/09/18 v1.11]*

Multiple imputation for the comparison of two screening tests in two-phase Alzheimer studies

Ofer Harel[†], X.H. (Andrew) Zhou[‡§]

[†] *Department of Statistics, University of Connecticut, 215 Glenbrook Road Unit 4120 Storrs, CT 06269-4120, USA*
[‡] *HSR&D Center of Excellence, VA Puget Sound Health Care System, 1660 South Columbian Way, 1/424, Seattle, WA 98108 USA*
[§] *Department of Biostatistics, School of Public Health, University of Washington, F600 Health Sciences, Box 357232, Seattle, WA 98195-7232, USA*

SUMMARY

Two-phase designs are common in epidemiological studies of dementia, and especially in Alzheimer research. In the first phase, all subjects are screened using a common screening test(s), while in the second phase, only a subset of these subjects is tested using a more definitive verification assessment, i.e. golden standard test. When comparing the accuracy of two screening tests in a two-phase study of dementia, inferences are commonly made using only the verified sample. It is well documented that in that case, there is a risk for bias, called verification bias. When the two screening tests have only two values (e.g. positive and negative) and we are trying to estimate the differences in sensitivities and specificities of the tests, one is actually estimating a confidence interval for differences of binomial proportions. Estimating this difference is not trivial even with complete data. It is well documented that it is a tricky task. In this paper, we suggest ways to apply imputation procedures in order to correct the verification bias. This procedure allows us to use well established complete-data methods to deal with the difficulty of the estimation of the difference of two binomial proportions in addition to dealing with incomplete data. We compare different methods of estimation, and evaluate the use of multiple imputation in this case. Our simulation results show that the use of multiple imputation is superior to other commonly used methods. We demonstrate our finding using an Alzheimer data.

1. Introduction

Two-phase designs are common in epidemiological studies of dementia [1, 2], and especially useful in Alzheimer research. In the first phase, all subjects are screened using a common screening test(s). While in the second phase, only a subset of these subjects is tested using a golden standard which is a clinical assessment. It is well documented that when only verified subjects are considered, there is a risk for bias, called verification bias [3]. There are several

*Correspondence to: Department of Statistics, University of Connecticut, 215 Glenbrook Road Unit 4120 Storrs, CT 06269-4120, USA

Contract/grant sponsor: Grant ; contract/grant number: 0101010101

Received 01 January 2005
Revised 01 January 2005

methods to deal with verification bias. The naive approach will use only verified subjects in the analysis. As it was stated above [3], this method can be biased. The most widely used correction method was developed by Begg and Greenes [4] under the ignorable verification bias assumption, which assumes the reason for selecting a sample for verification depends only on observed data. Zhou [5] extended that method using a maximum likelihood approach. Kosinski and Barnhart [6] suggested a method for correcting for non-ignorable verification bias. Zhou et al. [3] and Pepe [7] provided a good summary about this subject. An important question is how to compare the accuracies of two competing screening tests in discriminating between diseased and non-diseased subjects.

The same problem can be viewed as trying to estimate the confidence interval for the difference of two binomial proportions of paired data. Notice that two sensitivities and two specificities are both pairs of binomial proportions. The most used interval for the difference of paired binomial proportions is the Wald interval [8]. This interval has several limitations due to the nature of the Binomial distribution and the asymptotic theory on which it is based. Several alternative "exact" intervals have been proposed [9, p.123], [10], however due to the nature of the binomial distribution, Newcombe [11] showed that these intervals perform poorly as well. Newcombe [11] reviewed the literature and made comparisons using a simulation study of several methods. Based on his simulation study, he recommended a score interval with continuity correction called Newcombe hybrid (NH). An additional competing interval was studied by May and Johnson [12] (MJ). This interval was discussed by Lui [13], Newcombe [11], and Tango [14]. Zhou and Qin [15] proposed another confidence interval, based on the Edgeworth expansion. All these methods showed to be superior to the common Wald interval [8].

Estimating the sensitivity and specificity using the above mentioned procedures, requires complete data (i.e. both test results and true status for all subjects). Since this is rarely the case, it is useful to consider missing data procedures to deal with the incomplete data set. Harel and Zhou [16] showed that the use of missing data procedure performs better then the existing method when interested in the sensitivity and specificity of a test.

For example, our motivating example is an epidemiological study of dementia which investigate the role of environmental risk factors in the development of Alzheimer disease (AD). In this two-phase study, all participants are being screened using a two screening test, but only a portion of the sample undergos the golden standard for determining AD. Our goal is to compare between the two competing screening tests.

Multiple imputation (MI) [17] is a simulation based technique, replacing the missing values with m sets of plausible values, resulting in m sets of "complete" data sets. Computing the sensitivity and specificity estimates and their standard errors for each data set and combining them by simple arithmetic rules, gives a valid result taking into consideration the missing values. Using this method allows us not only to use the most common and simplest procedures to estimate the sensitivity and specificity, but also gives us grounds to compare different complete-data estimation procedures.

In the remainder of this article, we will set up the problem in section 2 and review the existing methods in section 3. In section 4 we will introduce the use of MI [17] to address the incomplete data sets issue, using methods for complete data mentioned in section 3. We will give a real data example in section 5 and compare the various techniques using a simple simulation study in section 6. We will discuss our conclusions in section 7.

Table I. Data Summary

		$T_1=1$		$T_1=0$	
		$T_2=1$	$T_2=0$	$T_2=1$	$T_2=0$
(a) Aggregated data					
$V=1$	$D=1$	x^A_{111}	x^A_{101}	x^A_{011}	x^A_{001}
	$D=0$	x^A_{110}	x^A_{100}	x^A_{010}	x^A_{000}
$V=0$		x^B_{11+}	x^B_{10+}	x^B_{01+}	x^B_{00+}
Total		n_{11+}	n_{10+}	n_{01+}	n_{00+}
(b) Complete data					
	$D=1$	x_{111}	x_{101}	x_{011}	x_{001}
	$D=0$	x_{110}	x_{100}	x_{010}	x_{000}
Total		n_{11+}	n_{10+}	n_{01+}	n_{00+}

2. Framework and data specification

Let T_1 and T_2 be two binary random variables, indicating whether or not the screening test was positive ($T_l=1$) or negative ($T_l=0$) where $l=1,2$. Since not all subjects' tests are being verified using the golden standard procedure, let V be a random variable indicating whether or not the subject was tested using the golden standard procedure ($V=1$ if tested, $V=0$ if not). Let D be the true disease status of a patient, such that, ($D=1$) if diseased, and ($D=0$) if non-diseased (we assume there is no measurement error for the golden standards procedure). Consider Table I(a) as a summary of aggregated representation for the data, where the $x's$ are the counts of observations in each status. One can consider $V=0$ to be the indication of missing data, since the test was not verified and we do not know the true status.

We can separate the data into two parts. First, when the screening tests, T_l, and the true status, D, are all observed ($V=1$), we can call it part A. Second, When T_l is observed but D is missing ($V=0$), we will refer to this as part B (Table I(a)). The total count of each cell x_{ijk} is a sum of two parts, $x_{ijk}=x^A_{ijk}+x^B_{ijk}$. Even-though x^A_{ijk} is totally observed, x^B_{ijk} is not, and instead we only observe the marginal total $x^B_{+jk}=x^B_{1jk}+x^B_{0jk}$. The observed data $Y_{obs}=\{x^A_{ijk},x^B_{+jk}:i,j,k=0,1\}$, are represented in Table I(a).

Consider the perfect scenario in which all subjects' test results were verified, and we have complete data (Table I(b)). Even in that case, estimating the specificity and sensitivity might not be a straightforward task. This estimation is the same as estimating the difference of two proportions from a binomial distribution. Newcombe [11] gave a detailed overview of this issue.

3. Existing methods

When true disease status is available for all subjects, the estimation of the sensitivity and specificity confidence intervals is the same as estimating the confidence interval of the difference of two binomial proportions. First, we review the methods for estimating these confidence intervals. Then we review the methods for estimating the confidence intervals in the incomplete

data case (i.e. true disease status is available only for a fraction of the subjects).

3.1. Complete data methods

3.1.1. McNemar's interval The common procedure to use when the parameter of interest is the difference of paired proportions is McNemar's interval (based on the McNemar's test) [18, pp.349-350]. Let (X_{0k}, X_{1k}), $k = 1, 2, \ldots, n$, be an independent and identically distributed sample from the joint distribution of the pair (X_0, X_1). Also, let X_0 and X_1 $(X_i, i = 1, 2)$ be correlated Bernoulli random variables with proportions p_1 and p_2, respectively. When our interest is in the difference $p = p_2 - p_1$, the Wald interval is based on the normal approximation of the distribution of the studentized difference between the two correlated sample proportions. The $100(1-\alpha)\%$ Wald interval is given by the following formula,

$$\hat{p} \pm z_{1-\alpha/2} n^{-1/2} \sqrt{\hat{p}_2(1-\hat{p}_2) + \hat{p}_1(1-\hat{p}_1) + 2(\hat{p}_1\hat{p}_2 - \hat{p}_{11})}, \tag{1}$$

where $Y_i = \Sigma_{k=1}^n X_{ik}$, $Y_{11} = \Sigma_{k=1}^n X_{0k}X_{1k}$, $\hat{p}_i = \frac{Y_i}{n}$, $\hat{p} = \hat{p}_2 - \hat{p}_1$, $\hat{p}_{11} = \frac{Y_{11}}{n}$, and z_α is the α-th quantile of the standard normal distribution.

3.1.2. McNemar's interval with continuity correction One of the problems with the binomial distribution is its lack of continuity. As a result, the normality assumption is violated. The continuity correction attempts to approximate the normal distribution more accurately. This correction brings the asymptotic distribution closer to a normal distribution [8, pp. 116-119]. In this case, the $100(1-\alpha)\%$ confidence interval will be

$$\hat{p} \pm (z_{1-\alpha/2} SE + 1/n), \tag{2}$$

where $\hat{p}$ is the difference of two binomial proportions estimate, SE is its standard error, z_α, is the α-th quantile of the standard normal distribution, and n is the sample size.

3.1.3. Newcombe hybrid (NH) Interval Newcombe [11] reviewed and compared several existing intervals for the difference between two binomial proportions based on paired data. Based on a simulation study, Newcombe [11] recommended a score interval with continuity correction called Newcombe hybrid (NH). In order to introduce this interval, let $Y_{00} = \Sigma_k(1 - X_{0k})(1 - X_{1k})$, $Y_{10} = \Sigma_k X_{0k}(1 - X_{1k})$, $Y_{01} = \Sigma_k(1 - X_{0k})X_{1k}$, and $D = (Y_{00} + Y_{10})(Y_{01} + Y_{11})(Y_{00} + Y_{01})(Y_{10} + Y_{11})$. Also let l_1 and u_1 be the lower and upper roots of the following quadratic equation of x: $(x - \frac{Y_{00}+Y_{01}}{n})^2 = (z_{1-\alpha/2})^2 \frac{x(1-x)}{n}$, and let l_2 and u_2 be the lower and upper roots of the following quadratic equation of x: $(x - \frac{Y_{00}+Y_{10}}{n})^2 = (z_{1-\alpha/2})^2 \frac{x(1-x)}{n}$. Therefore, the NH $100(1-\alpha)\%$ confidence interval is defined by:

$$[\hat{p} - (\delta_1^2 - 2\hat{\phi}\delta_1\epsilon_2 + \epsilon_2^2)^{1/2},\ \hat{p} - (\epsilon_1^2 - 2\hat{\phi}\epsilon_1\delta_2 + \delta_2^2)^{1/2}], \tag{3}$$

where

$$\begin{array}{ll} \delta_1 = \frac{Y_{00}+Y_{01}}{n} - l_1, & \epsilon_1 = u_1 - \frac{Y_{00}+Y_{01}}{n}, \\ \delta_2 = \frac{Y_{00}+Y_{10}}{n} - l_2, & \epsilon_2 = u_2 - \frac{Y_{00}+Y_{10}}{n}, \end{array}$$

and

$$\hat{\phi} = \begin{cases} \frac{Y_{00}Y_{11} - Y_{10}Y_{01}}{D} & Y_{00}Y_{11} - Y_{10}Y_{01} \le 0 \ \&\ D > 0 \\ \frac{Max(Y_{00}Y_{11} - Y_{10}Y_{01} - n/2, 0)}{D} & Y_{00}Y_{11} - Y_{10}Y_{01} > 0 \ \&\ D > 0 \\ 0 & D = 0 \end{cases}.$$

3.1.4. May and Johnson (MJ) interval Another competing interval was studied by May and Johnson [12] (MJ). This interval was discussed by Lui [13], Newcombe [11], and Tango [14]. In addition to the previous notation, let $A = (1 + \frac{z^2_{\alpha/2}}{n})$, $B = -2\frac{Y_{01}-Y_{10}}{n}$, and $C = (\frac{Y_{01}}{n} - \frac{Y_{10}}{n})^2 - z^2_{\alpha/2}\frac{Y_{01}+Y_{10}}{n^2}$. The MJ $(100-\alpha)\%$ confidence interval is:

$$\left[Max\{0, \frac{-B-(B^2-4AC)^{1/2}}{2A}\}, \;\; Min\{1, \frac{-B+(B^2-4AC)^{1/2}}{2A}\}\right]. \tag{4}$$

3.1.5. Zhou and Qin (ZQ) interval The validity of the Wald interval lies in the assumption that the data are normal. Since the true distribution of the Wald statistics is skewed, the normal approximation may produce bad results. Using the Edgeworth expansion, some of the bias might be corrected. To introduce this interval, we need some additional notation. Let

$$d = p_1(1-p_1)(1-2p_1) - p_2(1-p_2)(1-2p_2) + 6(p_1-p_2)(p_{11}-p_1p_2),$$

$\sigma = (p_1(1-p_1) + p_2(1-p_2) + 2(p_1p_2 - p_{11}))^{1/2}$, $a = \frac{d}{6\sigma^2}$, and $b = \frac{1-2p}{2} - \frac{d}{6\sigma^2}$,
where $p_{11} = P(X_0 = 1, X_1 = 1)$. Also, let us define a monotone transformation function, as $g(T) = \frac{\hat{a}\hat{\sigma}}{\sqrt{n}} + T + \frac{\hat{b}\hat{\sigma}T^2}{\sqrt{n}} + \frac{(\hat{b}\hat{\sigma})^2T^3}{4n}$ where $\hat{a}, \hat{b}, \hat{\sigma}, \hat{d}$ are the estimates of a, b, σ, d respectively. Using this transformation, the $100(1-\alpha)\%$ confidence interval for p, is given as follows:

$$\left[Max(-1, \hat{p} - \frac{\hat{\sigma}}{\sqrt{n}}g^{-1}(z_1 - \alpha/2)), Min(1, \hat{p} - \frac{\hat{\sigma}}{\sqrt{n}}g^{-1}(z_1 - \alpha/2))\right], \tag{5}$$

where

$$g^{-1}(y) = \begin{cases} \frac{\sqrt{n}}{\hat{b}\hat{\sigma}}\left[(1 + 3(\hat{b}\hat{\sigma})(\frac{y}{\sqrt{n}} - \frac{\hat{a}\hat{\sigma}}{n}))^{1/3} - 1\right] & \hat{b}\hat{\sigma} \neq 0 \\ y - \frac{\hat{a}\hat{\sigma}}{n} & \hat{b}\hat{\sigma} = 0 \end{cases}$$

3.2. Incomplete data methods

Oftentimes, not all subjects have been verified. Therefore, there is a chance for verification bias. When trying to compare the accuracy of two screening tests, the estimation of the differences of the sensitivities, and specificities is equivalent to estimating a difference of binomial proportions, but with incomplete data. Consider a sample of size n when we know that a sub-sample n_1 has a known true status D, while for $n_2 = n - n_1$ we do not have this information. This will require a different kind of methodology in order to estimate the unbiased population proportion.

3.2.1. Maximum Likelihood The only method available to date to deal with verification bias in paired comparisons of sensitivity and specificity was introduced by Zhou [19]. Under some ignorability conditions, the estimators for the sensitivities of two tests are as follows:

$$\hat{\pi}_{1ML} = \frac{\sum_{j=0}^{1}(x^A_{1j1}n_{1j+})/(x^A_{1j1} + x^A_{1j0})}{\sum_{i=0}^{1}\sum_{j=0}^{1}(x^A_{ij1}n_{ij+})(x^A_{ij1} + x^A_{ij0})},$$

and

$$\hat{\pi}_{2ML} = \frac{\sum_{i=0}^{1}(x^A_{i11}n_{i1+})/(x^A_{i11} + x^A_{i10})}{\sum_{i=0}^{1}\sum_{j=0}^{1}(x^A_{ij1}n_{ij+})(x^A_{ij1} + x^A_{ij0})}.$$

While the estimators for the specificities of two tests are:

$$\hat{\tau}_{1ML} = \frac{\sum_{j=0}^{1}(x^A_{0j0}n_{0j+})/(x^A_{0j0}+x^A_{0j1})}{\sum_{i=0}^{1}\sum_{j=0}^{1}(x^A_{ij0}n_{ij+})(x^A_{ij0}+x^A_{ij1})},$$

and

$$\hat{\tau}_{2ML} = \frac{\sum_{i=0}^{1}(x^A_{i00}n_{i0+})/(x^A_{i00}+x^A_{i01})}{\sum_{i=0}^{1}\sum_{j=0}^{1}(x^A_{ij0}n_{ij+})(x^A_{ij0}+x^A_{ij1})}.$$

In addition to the point estimates, Zhou [19] provided the estimates for the covariance matrices. The (k,l) element of the asymptotic covariance matrix of $\hat{\pi}_{1ML}$ and $\hat{\pi}_{2ML}$ is

$$\begin{aligned}\nu_{kl} &= \sum_{i,j=0}^{1}\frac{\theta^2_{ij}(1-\theta_{ij})^2}{x^A_{ij1}(1-\theta_{ij})^2+x^A_{ij0}\theta^2_{ij}}\frac{\partial\pi_k}{\partial\theta_{ij}}\frac{\partial\pi_l}{\partial\theta_{ij}}+\sum_{(i,j)\neq(1,1)}\frac{\eta^2_{ij}}{n_{ij}}\frac{\partial\pi_k}{\partial\eta_{ij}}\frac{\partial\pi_l}{\partial\eta_{ij}}\\ &\quad -\frac{1}{\sum_{i,j=0}^{1}\eta^2_{ij}/n_{ij}}\left(\sum_{(i,j)\neq(1,1)}\frac{\eta^2_{ij}}{n_{ij}}\frac{\partial\pi_k}{\partial\eta_{ij}}\right)\left(\sum_{(i,j)\neq(1,1)}\frac{\eta^2_{ij}}{n_{ij}}\frac{\partial\pi_l}{\partial\eta_{ij}}\right),\end{aligned}$$

and the (k,l) element of the asymptotic covariance matrix of $\hat{\tau}_{1ML}$ and $\hat{\tau}_{2ML}$ is

$$\begin{aligned}\nu_{kl} &= \sum_{i,j=0}^{1}\frac{\theta^2_{ij}(1-\theta_{ij})^2}{x^A_{ij1}(1-\theta_{ij})^2+x^A_{ij0}\theta^2_{ij}}\frac{\partial\tau_k}{\partial\theta_{ij}}\frac{\partial\tau_l}{\partial\theta_{ij}}+\sum_{(i,j)\neq(1,1)}\frac{\eta^2_{ij}}{n_{ij}}\frac{\partial\tau_k}{\partial\eta_{ij}}\frac{\partial\tau_l}{\partial\eta_{ij}}\\ &\quad -\frac{1}{\sum_{i,j=0}^{1}\eta^2_{ij}/n_{ij}}\left(\sum_{(i,j)\neq(1,1)}\frac{\eta^2_{ij}}{n_{ij}}\frac{\partial\tau_k}{\partial\eta_{ij}}\right)\left(\sum_{(i,j)\neq(1,1)}\frac{\eta^2_{ij}}{n_{ij}}\frac{\partial\tau_l}{\partial\eta_{ij}}\right),\end{aligned}$$

where $\theta_{ij} = P(D=1|T_1=i,T_2=j)$ and $\eta_{ij} = P(T_1=i,T_2=j)$ for $i,j=0,1$, and $\hat{\theta}_{ij} = \frac{x^A_{ij1}}{x^A_{ij1}+x^A_{ij0}}$, and $\hat{\eta}_{ij} = \frac{n_{ij}}{n}$. The $100(1-\alpha)\%$ confidence intervals for the differences of two sensitivities and specificities are

$$\hat{\pi}_2-\hat{\pi}_1 \pm z_{1-\alpha/2}\sqrt{\{\hat{Var}(\hat{\pi}_1)+\hat{Var}(\hat{\pi}_2)-2\hat{Cov}(\hat{\pi}_1,\hat{\pi}_2)\}}$$

and

$$\hat{\tau}_2-\hat{\tau}_1 \pm z_{1-\alpha/2}\sqrt{\{\hat{Var}(\hat{\tau}_1)+\hat{Var}(\hat{\tau}_2)-2\hat{Cov}(\hat{\tau}_1,\hat{\tau}_2)\}}$$

respectively.

4. MI for comparing two screening tests

Another method to deal with the verification bias when not all subjects have been verified is to use multiple imputation [16]. Multiple imputation (MI) [17, 20, 21] is a simulation technique to deal with missing data. We replace each missing value by $m>1$ plausible values, yielding m complete data sets that differ only in the imputed values. Analyzing each data set by a complete-data method described in Section 3.1 will result in m sets of point estimates and

standard errors. Combining the results by simple arithmetic rules will provide final estimates and standard errors taking into account the missing data.

In order for the MI to yield a valid inference, the simulated values must possess certain properties. MI drawn from a distribution with these qualities was called by Rubin [17] "proper". The full mathematical definition of proper MI is given by Rubin [17, pp.118–119]. Let Q and U be the population quantity of interest and its variance respectively, and let $\hat{Q}$ be its estimate. We assume that the data can be separated into X, all observed covariates, and $Y = (Y_{obs}, Y_{mis})$, observed and missing values. Since $\hat{Q}$ and $\bar{U}$ can be created using the imputed Y_{mis} together with the Y_{obs} and X, one needs the estimates from the imputed data sets to be unbiased for Q. For $j = 1, \ldots, m$ imputations, the large–m averages will be $E(\bar{Q}_\infty|X, Y) \doteq \hat{Q}$ and $E(\bar{U}_\infty|X, Y) \doteq U$ as m tends to infinity, while the between imputation variance will be $E(B_\infty|X, Y) \doteq Var(\bar{Q}_\infty|X, Y)$ for large m. Rubin [17] derives the procedure by Bayesian arguments. However, despite the Bayesian derivation, it has been shown that the method leads to inferences that are well calibrated from a frequentists standpoint [22, 23, 20, 21].

In addition, Meng [24] introduced the term congeniality. This term came to relate the Bayesian world and the frequentists' world. A model will be called uncongenial if the imputer model and the analysis model differ. A more mathematically rigorous definition is in Meng [24].

When the model is congenial and proper, we would get valid inference. If the model is not proper or uncongenial, we will get valid inference only part of the time, depending on the specific scenario. In the next section, we propose a proper MI procedure for correcting for verification bias.

4.1. Imputation stage

The main step of MI is to derive the posterior distribution of the missing disease statuses, given their test results (either positive or negative). Throughout the imputation procedure, we use data augmentation [25] for imputing the missing values. Under the ignorability assumption and the structure of the data in Table I(a), one can look at the data as if it came from a multinomial distribution. We can use the multinomial property, in which a conditional multinomial is a multinomial as well (see Appendix 1), to derive the predictive distribution of missing data given the observed data, which is given as follows:

$$(x^B_{1jk}, x^B_{0jk})|Y_{obs}, \theta \sim M(x^B_{+jk}, (\theta_{1jk}/\theta_{+jk}, \theta_{0jk}/\theta_{+jk})), \;\; j, k = 0, 1,$$

where θ_{ijk} is the probability that a unit falls into cell (i, j, k), $\theta_{+jk} = \sum_i \theta_{ijk}$, and $M(.,.)$ represents a multinomial distribution. By indexing the cells in the contingence table using only one subscript $(d = 1, ..., D)$, it follows that

$$x|\theta \sim M(n, \theta)$$

with $\theta = (\theta_1, \theta_2, ..., \theta_D)$,

When choosing a Dirichlet prior distribution for multinomial parameters, we obtain the following results which are well known from the conjugate family idea in Bayesian statistics (see Appendix 2).

$$x|\theta \sim M(n, \theta), \tag{6}$$

$$\theta \sim D(\alpha), \tag{7}$$

$$\theta|Y \sim D(\alpha'), \tag{8}$$

where $\alpha' = \alpha + x$, and $D(\alpha)$ is a Dirichlet distribution with parameter α.

The data augmentation procedure is drawing iteratively from two distributions. First, one should draw the $x's$ from a multinomial distribution (6). This is done under the assumption that θ is known. Then given those $x's$ values, one should draw values for θ from the (Dirichlet, Beta) posterior distribution (8). The imputation can be carried forward easily using any MI software which allows categorical or loglinear models. For example, Splus [26]. The computational details can be found in Schafer [21].

The scheme for the imputation stage follows proper imputation draws. Schafer [21] elaborates on the properties of this model. The use of Jeffreys prior is a common practice in Bayesian analysis when one wants to use a non informative prior [27].

4.2. Analysis stage

After imputing the missing-data statuses, we obtain m sets of complete data sets. Using complete-data methods outlined in Section 3.1 we obtain the estimates $(\hat{Q}^{(1)}, \hat{Q}^{(2)}, \ldots, \hat{Q}^{(m)})$ and associated variances $(U^{(1)}, U^{(2)}, \ldots, U^{(m)})$ for the differences of the sensitivities and specificities. The complete-data methods we are going to use are: McNemar's interval (McN)[11], section 3.1.1; McNemar's interval with continuity correction (McN+CC) [11, Method #2], [8, pp. 116-119], section 3.1.2; Newcombe-Hybrid (NH) [11], section 3.1.3; May-Johnson (MJ) [12], section 3.1.4; Zhou-Qin (ZQ) [15], section 3.1.5.

4.3. Combining results

After having m sets of estimates and variances, we use Rubin's [17] combining rules as follows: The overall estimate is $\bar{Q} = \frac{1}{m}\sum Q^{(i)},\ i = 1, \ldots, m$, and its variance is $T = \bar{U} + \frac{1}{m+1}B$, where $\bar{U} = \frac{1}{m}\sum U^{(i)}$ is the complete-data variance estimate, and $\frac{1}{m+1}B$ is the variance addition due to the imputations of missing values. The inferences are based on the t-distribution approximation $T^{-1/2}(Q-\hat{Q}) \sim t_\nu$ where the degrees of freedom are $\nu = (m-1)[1+\frac{\bar{U}}{(1+m^{-1})B}]^2$. Therefore, the $100(1-\alpha)\%$ confidence interval for the estimate will be $\bar{Q} \pm t_{\nu,1-\alpha/2}\sqrt{T}$.

5. Data Example

In this section we introduce the motivating example from Alzheimer research.

5.1. Environmental risk factors for the development of Alzheimer's disease

The first motivating example is an epidemiological study of dementia which investigated the role of environmental risk factors in the development of Alzheimer's disease [28]. One of the aims of the study was to compare the existing (standard) screening test to a new one. The new test [29] is based on information from a cognitive test given to a person and from a relative test given to someone who knows the subject. The standard test uses only the information from the subject's test [30]. The results for older adults are summarized in Table II.

Following the notation of section 2, our data can be represented as below, which include the observed data and the aggregated data which contains the missing information due to missing

Table II. Data from an Alzheimer study ($Age > 75$)

	Test 1 positive		Test 1 negative	
	Test 2 positive	Test 2 negative	Test 2 positive	Test 2 negative
Verified				
Disease	31	5	3	1
Non-disease	25	10	19	55
Not verified	22	6	65	346
Total	78	21	87	402

values

$$\begin{aligned} Y_{obs} &= \{x^A_{111}=31, x^A_{101}=5, x^A_{011}=3, x^A_{001}=1, \\ &\quad x^A_{110}=25, x^A_{100}=10, x^A_{010}=19, x^A_{000}=55, \\ &\quad x^B_{11+}=22, x^B_{10+}=6, x^B_{01+}=65, x^B_{00+}=346\}. \end{aligned}$$

In order to proceed with the data augmentation algorithm, let us choose the parameter for the prior Dirichlet distribution to be
$\alpha = (1.5, 1.5, 1.5, 1.5, 1.5, 1.5, 1.5, 1.5)$, which implies Jeffreys prior. Therefore, our predictive distributions are as follows:

$$\begin{aligned} (x^B_{110}, x^B_{111})^{t+1}|Y_{obs}, \theta^t &\sim M(x^B_{11+}, (\theta_{111}/\theta_{11+}, \theta_{110}/\theta_{11+})) \\ (x^B_{100}, x^B_{101})^{t+1}|Y_{obs}, \theta^t &\sim M(x^B_{10+}, (\theta_{101}/\theta_{10+}, \theta_{100}/\theta_{10+})) \\ (x^B_{010}, x^B_{011})^{t+1}|Y_{obs}, \theta^t &\sim M(x^B_{01+}, (\theta_{011}/\theta_{01+}, \theta_{010}/\theta_{01+})) \\ (x^B_{000}, x^B_{001})^{t+1}|Y_{obs}, \theta^t &\sim M(x^B_{00+}, (\theta_{001}/\theta_{00+}, \theta_{000}/\theta_{00+})). \end{aligned}$$

The prior is $\theta \sim D(\alpha)$ and the posterior is

$$\theta|Y \sim D(x_{111}+0.5, x_{110}+0.5, x_{101}+0.5, x_{100}+0.5, x_{011}+0.5, x_{010}+0.5, x_{001}+0.5, x_{000}+0.5),$$

where $x_{ijk} = x^A_{ijk} + x^B_{ijk}$ $i, j, k = 0, 1$, and t is the number of iterations. Using S-plus 6.2 [26] we use MI ($m = 10$) to compare the existing methods described in section 3.2, with the methods described in section 3.1. Table III summarizes the results, where est represent the appropriate estimate for the differences of sensitivities or specificities, with SE as its standard error, and up, low are the upper and lower levels of the confidence interval. We try to estimate the differences of sensitivities and specificities of Test 1 (new) and Test 2 (existing method), such that $est = est_1 - est_2$

It follows that all the MI sensitivity estimates are quite close to each other, differences of only thousandths have been found. The ML estimate is smaller, but still very close to the other estimates (less then a half of SE). The confidence intervals are similar with the exception of the MJCI, which results in different inferential results. (There is significant difference between the two tests). The specificity estimates are all close to each other, differences of only thousandths have been found. It seems that all tests find significant differences between the two tests except the NHCI which claims no difference, and the MJCI which cannot be estimated. We can conclude that there is no difference between the sensitivities of the two tests, but the new test improves the specificity.

Table III. Results comparing MI methods, the ML as an existing method of the differences of two sensitivities and specificities – Alzheimer data

		Multiple Imputation					
		NHCI	MJCI	ZQCI	McN	McN+CC	ML
Sensitivity	est	0.1082	0.1054	0.1082	0.1082	0.1082	0.0703
	SE				0.0928	0.0928	0.0929
	up	0.9542	0.1870	0.2195	0.2982	0.3121	0.2524
	low	-0.6600	0.0433	-0.0032	-0.0819	-0.0957	-0.1118
Specificity	est	-0.1122	-0.1118	-0.1122	-0.1122	-0.1122	-0.1178
	SE				0.0201	0.0201	0.0201
	up	0.3325		-0.0804	-0.0725	-0.0706	-0.0785
	low	-0.5993		-0.1440	-0.1519	-0.1538	-0.1572

6. Simulations

In order to compare the different estimation methods and evaluate the use of MI to correct for verification bias, we run two sets of simulation studies. We compare the estimates for the differences of sensitivities and specificities in terms of bias, mean square error (MSE), the corresponding confidence intervals in terms of interval length and true coverage. The first set of simulations are based on the Alzheimer data set in section 5.1. We chose this example since it was published already, and there are several cells with very few subjects in them. For the second set of simulations we used some of the settings from the first simulations but we have increased the sample size. The settings of the simulation studies are as follows: sample size of $N = (588, 1000)$, prevalence $p = P(D = 1) = 0.35$, $\lambda_{11} = \lambda_{10} = 0.7$, $\lambda_{01} = 0.25$ and $\lambda_{00} = 0.14$, where $\lambda_{ij} = P(V = 1|T_1 = i, T_2 = j)$. The sensitivity $Se_k = P(T_k = 1|D = 1)$, and specificity $Sp_l = P(T_l = 0|D = 0)$ are stated in the tables.

We run this simulation 10,000 times. For our MI procedures we take $m = 10$, using S-plus 6.2 [26] with a flat (noninformative) prior. The results are summarized in Tables IV–VI.

Let us consider the results summarized in Table IV(a). It follows that all the sensitivity and specificity estimates are quite close to each other, differences of only thousandths have been found. The ML estimate is a bit different than the other estimates. But as for the MSE it is almost the same, and with this size of difference, we can assume they are the same. We can see quite a variation in the CI length, but together with that there is the issue of coverage. It follows that the McN, McN+CC and ML are distinctly different than the other methods. However, less has to be said about the quality of the CI. In this simulation we are showing that the multiple imputation with McN, McN with continuity correction, and the maximum likelihood have very similar results, while other methods do not perform that well. We can see that the coverage which is supposed to be 95% is very close to that in the McN, McN+CC, and ML, but quite far for the NH, and MJ. ZQ is positioned somewhere between the two groups with respect to coverage.

The results in Table V(a) are a bit different. Notice that although the estimates for the sensitivity and specificities are all very close and the biases and MSE are similar as well, the coverage of the ML method is pretty bad (approx 50%) while the other methods are the same as the previous tables. Due to this big change, the ZQ coverage is much better than the ML

coverage in these settings. The results in Table VI(a), again, are very similar with regard to all the equality measures, except that in this case the ML and ZQ coverage are very close to each other, while the other methods are quite constant in their coverage levels.

When considering the simulations for bigger data sets, the results are very similar. In table IV(b) the estimates, biases, and MSE are very close to one another and to what we would expect. When considering the coverage, again, Table IV(b) show that the coverage that is supposed to be 95% is very close to that in the McN, McN+CC, and ML, but quite far for the NH, and MJ. The ZQ is positioned somewhere between the two groups with respect to coverage. Under the larger sample size simulations, Table VI(b) has similar results, and only in Table V(b) does the ML coverage perform badly.

From the simulation study we learn that the use of MI procedure with McNemera's interval with continuity correction is the recommended method. There is only a limited difference between the simple McNemera's interval to the one with continuity correction, but in most cases the continuity correction give better results (See Tables IV–VI). Also when using MI it is very simple to use more than one analysis procedure which will allow us to do sensitivity analysis. It seems that when the estimates are getting closer to the boundary, the ML method do not perform as well (Table V). Also when the sample size is small and the difference in sensitivities and specificities are different, the ML does not perform as well as the MI.

7. Discussion

We have proposed the use of MI for comparing the accuracies of two screening tests in a two-phase study design. Two-phase study designs are very popular. In particular, our motivation came from Alzheimer research in which only a portion of a study sample can be verified (using clinical assessment). From a theoretical point of view, comparing the accuracy of two screening tests in a two-phase study design is equivalent to estimating the confidence interval of the difference of two binomial proportions of paired data. Both theoretical and practical problems are of interest for researchers.

The use of MI allows us to use several complete-data analysis methods in the analysis stage. This advantage permits us to use sensitivity analysis and compare the different complete-data methods. In addition, the cost of using several methods is nominal, as the code is exactly the same except in the analysis stage. Simulation studies have been used in order to compare the different complete-data procedures between themselves, and also to compare the MI procedure in general (choosing one complete-data method) to the existing incomplete-data procedure (maximum likelihood). Comparing the complete-data methods, it is quite apparent that the use of MI with the McNemar's (with or without continuity correction), is the best method. Although there are not many differences in estimation, bias, and mean squared errors, the coverage and confidence interval length are much better for the McNemar's test comparing to NH, MJ, and ZQ. In most cases the continuity correction helps the coverage, and therefore our method of choice (among the MI methods) will be the McNemar's interval with continuity correction. When comparing the MI McNemar's intervals to the existing method (ML), we see that in most cases the estimates, bias, and mean squared errors are quite comparable. On the other hand, the difference in coverage can be quit high. Therefore, we concluded that our

Table IV. Simulation results (estimate, Bias, MSE, lower and upper levels, CI length and Coverage) for the difference of two screening tests ($t_2 - t_1$). True values are $Se_1 = Se_2 = 0.9$, $Sp_1 = Sp_2 = 0.9$, and 95% coverage

(a) N=588	Sensitivity						
	Est	Bias	MSE	lower	upper	length	Coverage
McN	0.0026	0.0026	0.0019	-0.0849	0.0901	0.1749	92.6
McN+CC	0.0026	0.0026	0.0019	-0.0899	0.0951	0.1849	94.5
NH	0.0026	0.0026	0.0019	-0.3601	0.3652	0.7253	100.0
MJ	0.0026	0.0026	0.0019	0.0097	0.5055	0.4958	23.1
ZQ	0.0026	0.0026	0.0019	-0.0544	0.0595	0.1139	80.3
ML	0.0007	0.0007	0.0020	-0.0854	0.0868	0.1722	95.0
	Specificity						
	Est	Bias	MSE	lower	upper	length	Coverage
McN	0.0004	0.0004	0.0008	-0.0536	0.0544	0.1079	92.6
McN+CC	0.0004	0.0004	0.0008	-0.0562	0.0570	0.1131	94.5
NH	0.0004	0.0004	0.0008	-0.3496	0.3511	0.7006	100.0
MJ	0.0004	0.0004	0.0008	0.0060	0.5108	0.5048	28.6
ZQ	0.0004	0.0004	0.0008	-0.0406	0.0414	0.0819	84.9
ML	-0.0010	-0.0010	0.0008	-0.0545	0.0524	0.1069	93.8

(b) N=1000	Sensitivity						
	Est	Bias	MSE	lower	upper	length	Coverage
McN	-0.0006	-0.0006	0.0011	-0.0668	0.0656	0.1325	94.0
McN+CC	-0.0006	-0.0006	0.0011	-0.0697	0.0685	0.1381	95.7
NH	-0.0006	-0.0006	0.0011	-0.3252	0.3225	0.6477	100
MJ	-0.0006	-0.0006	0.0011	0.0277	0.0699	0.0422	0
ZQ	-0.0006	-0.0006	0.0011	-0.0434	0.0422	0.0855	79.7
ML	-0.0017	-0.0017	0.0011	-0.0641	0.0610	0.1250	92.4
	Specificity						
	Est	Bias	MSE	lower	upper	length	Coverage
McN	0.0001	0.0001	0.0005	-0.0424	0.0425	0.0848	94.8
McN+CC	0.0001	0.0001	0.0005	-0.0439	0.0440	0.0879	95.6
NH	0.0001	0.0001	0.0005	-0.2924	0.2928	0.5852	100
MJ	0.0001	0.0001	0.0005	0.0159	0.0404	0.0244	0
ZQ	0.0001	0.0001	0.0005	-0.0320	0.0321	0.0641	88.6
ML	-0.0008	-0.0008	0.0005	-0.0420	0.0406	0.0827	93.3

Table V. Simulation results (estimate, Bias, MSE, lower and upper levels, CI length and Coverage) for the difference of two screening tests ($t_2 - t_1$). True values are $Se_1 = Se_2 = 0.9$, $Sp_1 = Sp_2 = 0.95$, and 95% coverage

(a) N=588	Sensitivity						
	Est	Bias	MSE	lower	upper	length	Coverage
McN	-0.0012	-0.0012	0.0016	-0.0792	0.0767	0.1559	93.0
McN+CC	-0.0012	-0.0012	0.0016	-0.0840	0.0816	0.1656	95.0
NH	-0.0012	-0.0012	0.0016	-0.3342	0.3293	0.6635	100
MJ	-0.0012	-0.0012	0.0016	0.0252	0.0841	0.0589	0
ZQ	-0.0012	-0.0012	0.0016	-0.0561	0.0536	0.1096	81.3
ML	-0.0013	-0.0013	0.0018	-0.0818	0.0769	0.1587	49.7
	Specificity						
	Est	Bias	MSE	lower	upper	length	Coverage
McN	0.0004	0.0004	0.0005	-0.0421	0.0429	0.0850	93.6
McN+CC	0.0004	0.0004	0.0005	-0.0447	0.0456	0.0903	95.0
NH	0.0004	0.0004	0.0005	-0.1584	0.1605	0.3189	100
MJ	0.0004	0.0004	0.0005	0.0130	0.0456	0.0325	0
ZQ	0.0004	0.0004	0.0005	-0.0299	0.0307	0.0606	85.5
ML	0.0000	0.0000	0.0005	-0.0449	0.0408	0.0857	48.3

(b) N=1000	Sensitivity						
	Est	Bias	MSE	lower	upper	length	Coverage
McN	-0.0016	-0.0016	0.0010	-0.0626	0.0594	0.1220	94.1
McN+CC	-0.0016	-0.0016	0.0010	-0.0654	0.0623	0.1277	95.1
NH	-0.0016	-0.0016	0.0010	-0.3194	0.3145	0.6339	100
MJ	-0.0015	-0.0015	0.0010	0.0243	0.0654	0.0412	0
ZQ	-0.0016	-0.0016	0.0010	-0.0444	0.0412	0.0856	83.1
ML	-0.0014	-0.0014	0.0010	-0.0605	0.0571	0.1176	63.7
	Specificity						
	Est	Bias	MSE	lower	upper	length	Coverage
McN	0.0006	0.0006	0.0003	-0.0324	0.0337	0.0661	94.9
McN+CC	0.0006	0.0006	0.0003	-0.0340	0.0353	0.0692	95.8
NH	0.0006	0.0006	0.0003	-0.1512	0.1535	0.3047	100
MJ	0.0006	0.0006	0.0003	0.0116	0.0333	0.0217	0
ZQ	0.0006	0.0006	0.0003	-0.0227	0.0240	0.0468	84.2
ML	0.0006	0.0006	0.0003	-0.0323	0.0322	0.0645	64.3

Table VI. Simulation results (estimate, Bias, MSE, lower and upper levels, CI length and Coverage) for the difference of two screening tests ($t_2 - t_1$). True values are $Se_1 = 0.8$, $Se_2 = 0.9$, $Sp_1 = Sp_2 = 0.9$, and 95% coverage

(a) N=588	Sensitivity						
	Est	Bias	MSE	lower	upper	length	Coverage
McN	0.0965	-0.0035	0.0024	-0.0008	0.1938	0.1946	94.0
McN+CC	0.0965	-0.0035	0.0024	-0.0056	0.1986	0.2042	95.2
NH	0.0965	-0.0035	0.0024	-0.3774	0.6240	1.0014	99.9
MJ	0.0956	-0.0044	0.0024	0.0719	0.1566	0.0847	58.3
ZQ	0.0965	-0.0035	0.0024	0.0314	0.1616	0.1302	80.3
ML	0.0991	-0.0009	0.0025	0.0009	0.1959	0.1950	87.8
	Specificity						
	Est	Bias	MSE	lower	upper	length	Coverage
McN	-0.0001	-0.0001	0.0009	-0.0595	0.0593	0.1188	93.9
McN+CC	-0.0001	-0.0001	0.0009	-0.0621	0.0619	0.1240	94.6
NH	-0.0001	-0.0001	0.0009	-0.3017	0.3024	0.6040	100
MJ	-0.0001	-0.0001	0.0009	0.0217	0.0603	0.0385	0
ZQ	-0.0001	-0.0001	0.0009	-0.0417	0.0415	0.0832	82.6
ML	0.0000	0.0000	0.0010	-0.0579	0.0575	0.1155	86.6

(b) N=1000	Sensitivity						
	Est	Bias	MSE	lower	upper	length	Coverage
McN	0.0973	-0.0027	0.0014	0.0213	0.1732	0.1519	95.4
McN+CC	0.0973	-0.0027	0.0014	0.0185	0.1761	0.1576	96.5
NH	0.0973	-0.0027	0.0014	-0.3600	0.5975	0.9575	100
MJ	0.0967	-0.0033	0.0014	0.0727	0.1350	0.0624	59.5
ZQ	0.0973	-0.0027	0.0014	0.0467	0.1478	0.1012	81.1
ML	0.0989	-0.0011	0.0014	0.0264	0.1716	0.1452	93.5
	Specificity						
	Est	Bias	MSE	lower	upper	length	Coverage
McN	0.0006	0.0006	0.0006	-0.0456	0.0468	0.0924	95.1
McN+CC	0.0006	0.0006	0.0006	-0.0472	0.0484	0.0955	95.4
NH	0.0006	0.0006	0.0006	-0.2923	0.2942	0.5865	100
MJ	0.0006	0.0006	0.0006	0.0200	0.0465	0.0265	0
ZQ	0.0006	0.0006	0.0006	-0.0315	0.0327	0.0641	83.1
ML	0.0006	0.0006	0.0005	-0.0438	0.0448	0.0886	93.3

proposed method performs better than the existing method (at least with respect to nominal coverage).

One of the major assumptions made is the ignorability assumption. One can argue whether this assumption is reasonable or not, but since both MI and ML assume the same assumption, we can say that the comparison is just. In the case in which the ignorability assumption is not reasonable, one can still use the MI method, but the use of ML is questionable. When using MI under non-ignorable missingness, one would need to change the imputation model. That implies that the missing values will be generated from a different distribution $P(Y_{mis}|Y_{obs}, R)$ instead of $P(Y_{mis}|Y_{obs})$, where R can be considered a random variable that separates the data into the observed and missing parts. Other than that, all the procedure will be the same.

In our manuscript we compare two screening tests in an Alzheimer two-phase design study. In some cases the two levels of the study are not an actual two-phase sampling design, as the second level sampling units are sampling according to a process outside the researchers control. For example, In AD research many research questions can be answered only by autopsy. The researchers (in advance) can not specify who is going to undertake autopsy and who is not. Our method will still be valid for the case of not conventional two-phase designs.

Another complication that can affect the validity of the ignorability assumption, is the fact that in the background there are two processes that determine who will undertake autopsy, and who will not. First, a subject has to die in order to be considered for autopsy. In this case we know, that those who are still alive are missing at random. Second, those who died would either be autopsied ar not, and this process is different then the first one and can be considered as ignorable or not. The solution for this complication is the subject of a following manuscript.

Our example (section 5.1) tried to find if there are environmental risk factors for the development of Alzheimer's disease. The standard screening test was based on the subject test, while the new test was based on the subject's test in addition to information from a relative test given to someone who knows the subject. We found (again) that there are no differences in the sensitivities of the two tests, but that there is an improvement in specificity. This result support the results of Zhou [19] when only the ML method is used.

Appendix 1 – Multinomial properties

Let x be a multinomial random variable with parameter θ. By indexing the cells in the contingence table using only one subscript ($d = 1, ..., D$), it follows that

$$x|\theta \sim M(n, \theta)$$

with $\theta = (\theta_1, \theta_2, ..., \theta_D)$, where the probability distribution of x is

$$P(x|\theta) = \frac{n!}{x_1!x_2!\cdots x_D!}\theta_1^{x_1}\theta_2^{x_2}\cdots\theta_D^{x_D}$$

Suppose that we collapse two cells of the contingency table, adding the frequencies together, such that we produce a new table $x^* = (z, x_3, ..., x_D)$, where $z = x_1 + x_2$.

Result 1. *The distribution of x^* is multinomial such that*

$$x^*|\theta \sim M(n, \theta^*),$$

where $\theta^* = (\xi, \theta_3, ..., \theta_D)$, *and* $\xi = \theta_1 + \theta_2$.

Proof 1. Let us sum the multinomial probabilities for all the x-vectors consistent with z, such that

$$\begin{aligned} P(x^*|\theta) &= \sum_{j=0}^{z} P(x_1 = j, x_2 = z - j, x_3, ..., x_D) \\ &= \sum_{j=0}^{z} \frac{n!}{j!(z-j)!x_3! \cdots x_D!} \theta_1^j \theta_2^{z-j} \theta_3^{x_3} \cdots \theta_D^{x_D} \\ &= \frac{n!}{z!x_3! \cdots x_D!} \theta_3^{x_3} \cdots \theta_D^{x_D} \sum_{j=0}^{z} \frac{z!}{j!(z-j)!} \theta_1^j \theta_2^{z-j} \\ &= \frac{n!}{z!x_3! \cdots x_D!} \theta_3^{x_3} \cdots \theta_D^{x_D} (\theta_1 + \theta_2)^z \end{aligned}$$

since $\sum_{j=0}^{z} \frac{z!}{j!(z-j)!} \theta_1^j \theta_2^{z-j} = (\theta_1 + \theta_2)^z$.

Result 2. *The conditional distribution of* (x_1, x_2) *given* z *(the sum) is multinomial such that*

$$(x_1, x_2)|z, \theta \sim M(z, (\theta_1/\xi, \theta_2/\xi)).$$

Proof 2. By using result 1 continuously on variables x_3 to x_D, those cells will collapse to a single cell such that $x_3 + \cdots + x_D = n - z$. Therefore,

$$\begin{aligned} (x_1, x_2, n - z)|\theta &\sim M(n, (\theta_1, \theta_2, 1 - \xi)) \\ (z, n - z)|\theta &\sim M(n, (\xi, 1 - \xi)). \end{aligned}$$

By the definition of conditional probability, it follows that

$$P(x_1, x_2|z, \theta) = \frac{P(x_1, x_2, z|\theta)}{P(z, n - z|\theta)} = \frac{P(x_1, x_2, n - z|\theta)}{P(z, n - z|\theta)},$$

Since both numerator and denominator are multinomial distributions, we can replace the expressions on the right hand side to get

$$\left[\frac{n!}{x_1!x_2!(n-z)!} \theta_1^{x_1} \theta_2^{x_2} (1 - \xi)^{n-z} \right] \left[\frac{n!}{z!(n-z)!} \xi^z (1 - \xi)^{n-z} \right]^{-1}$$

which can be reduced to $P(x_1, x_2|z, \theta) = \frac{z!}{x_1!,x_2!} (\theta_1/\xi)^{x_1} (\theta_2/\xi)^{x_2}$, the desired result.

Although the results are stated such that the collapsing is of two cells, the results are true for any arbitrary sets of collapsing.

Appendix 2 – Dirichlet prior

Let $\theta = (\theta_1, \theta_2, ..., \theta_D)$ be a set of random variables such that $\theta_d \geq 0$ for $d = 1, 2, ..., D$ and $\sum_{d=1}^{D} \theta_d = 1$. The density function of θ given the parameter $\alpha = (\alpha_1, \alpha_2, ...\alpha_D)$, is

$$P(\theta|\alpha) = \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)\Gamma(\alpha_2) \cdots \Gamma(\alpha_D)} \theta_1^{\alpha_1 - 1} \theta_2^{\alpha_2 - 1} \cdots \theta_D^{\alpha_D - 1}$$

where $\alpha_0 = \sum_{d=1}^{D} \alpha_d$ and $\Gamma(.)$ denotes the gamma function. This Dirichlet distribution is often written as $\theta|\alpha \sim D(\alpha)$. When used as a prior for a multinomial distribution, it is typical to omit the normalizing constant such that,

$$\pi(\theta) \propto \theta_1^{\alpha_1 - 1}\theta_2^{\alpha_2 - 1} \cdots \theta_D^{\alpha_D - 1}$$

where $(\alpha_1, ..., \alpha_D)$ are user specific hyperparameters. Since the likelihood function of a multinomial distribution is

$$L_{x|\theta} = \frac{n!}{x_1 x_2! \cdots x_D!}\theta_1^{x_1}\theta_2^{x_2} \cdots \theta_D^{x_D},$$

the posterior distribution is the product of the prior function (information) and the likelihood function, leading us to

$$\begin{aligned} L_{\theta|x} = \pi(\theta) \times L_{x|\theta} &\propto K \times (\theta_1^{\alpha_1 - 1}\theta_2^{\alpha_2 - 1} \cdots \theta_D^{\alpha_D - 1})(\theta_1^{x_1}\theta_2^{x_2} \cdots \theta_D^{x_D}) \\ &= K \times \theta_1^{x_1 + \alpha_1 - 1}\theta_2^{x_2 + \alpha_2 - 1} \cdots \theta_D^{x_D + \alpha_D - 1} \\ &\sim D(x + \alpha), \end{aligned}$$

a Dirichlet posterior distribution with parameter $(x + \alpha) = (x_1 + \alpha_1, x_2 + \alpha_2, ..., x_D + \alpha_D)$.

references

[1] D. J. Hand. Screening vs prevalence estimation. *Applied Statistics*, 36:1–7, 1987.

[2] Andrew Pickles, Graham Dunn, and Jos Luis Vzquez-Barquero. Screening for stratification in two-phase ('two-stage') epidemiological surveys. *Statistical Methods in Medical Research*, 4:73–89, 1995.

[3] X.H. Zhou, N.A. Obuchowski, and D.M. Obuchowski. *Statistical Methods is Diagnostic Medicine*. Wiley & Sons, New York, USA, 2002.

[4] Colin B. Begg and Robert A. Greenes. Assessment of diagnostic tests when disease verification is subject to selection bias. *Biometrics*, 39:207–215, 1983.

[5] Xiao-Hua Zhou. Maximum likelihood estimators of sensitivity and specificity corrected for verification bias. *Communications in Statistics, Part A – Theory and Methods [Split from: @J(CommStat)]*, 22:3177–3198, 1993.

[6] A S Kosinski and H X Barnhart. Accounting for nonignorable verification bias in assessment of diagnostic tests. *Biometrics*, 59:163–171, March 2003.

[7] M S Pepe. *The Statistical Evaluation of Medical Tests of Classification and Prediction*. Oxford University Press, 2003.

[8] Joseph L. Fleiss. *Statistical methods for rates and proportions*. John Wiley & Sons, 1981.

[9] F.D.K. Liddell. Simplified exact analysisof case-referent studies: matched pairs; dichotomous outcome. *Journal of Epidemiology and Community Health*, 37:82–84, 1983.

[10] P. Armitage and G. Berry. *Statistical methods in medical research (Second edition)*. Blackwell Scientific Publications Ltd, 1987.

[11] Robert G. Newcombe. Improved confidence intervals for the difference between binomial proportions based on paired data. *Statistics in Medicine*, 17:2635–2650, 1998.

[12] Warren L. May and William D. Johnson. Confidence intervals for differences in correlated binary proportions (Corr: 1998V17 p2015). *Statistics in Medicine*, 16:2127–2136, 1997.

[13] Kung-Jong Lui. Comment on "Confidence intervals for differences in correlated binary proportions" (1997V16 p2127-2136). *Statistics in Medicine*, 17:2017–2020, 1998.

[14] Toshiro Tango. Equivalence test and confidence interval for the difference in proportions for the paired-sample design. *Statistics in Medicine*, 17:891–908, 1998.

[15] X.H. Zhou and G.S. Qin. A new confidence interval for the difference between two binomial proportions of paired data. *Journal of Statistical Planning and Inference*, 128:527–542, 2005.

[16] O. Harel and X.H. Zhou. Multiple imputation for correcting verification bias. *Statistics in Medicine*, 2005. In Press.

[17] D B Rubin. *Multiple Imputation for Nonresponse in Surveys*. J. Wiley & Sons, New York, 1987.

[18] Alan Agresti. *Categorical data analysis*. John Wiley & Sons, 2002.

[19] Xiao-Hua Zhou. Comparing accuracies of two screening tests in a two-phase study for dementia. *Applied Statistics*, 47:135–147, 1998.

[20] D B Rubin. Multiple imputation after 18+ years. *Journal of American Statistical Association*, 91:473–489, 1996.

[21] J L Schafer. *Analysis of Incomplete Multivariate Data*. Chapman & Hall, London, 1997.

[22] Donald B. Rubin and Nathaniel Schenker. Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. *Journal of the American Statistical Association*, 81:366–374, 1986.

[23] Nathaniel Schenker and A. H. Welsh. Asymptotic results for multiple imputation. *The Annals of Statistics*, 16:1550–1566, 1988.

[24] Xiao-Li Meng. Multiple-imputation inferences with uncongenial sources of input (Disc: p558-573). *Statistical Science*, 9:538–558, 1994.

[25] Martin A. Tanner and Wing Hung Wong. The calculation of posterior distributions by data augmentation (C/R: p541-550). *Journal of the American Statistical Association*, 82:528–540, 1987.

[26] J Schimert, J L Schafer, T Hesterberg, C Fraley, and D Clarkson. *Analyzing Missing Values in S-PLUS*. Insightful Corp., Seattle, WA, 2001.

[27] Robert E. Kass and Larry Wasserman. The selection of prior distributions by formal rules (Corr: 1998V93 p412). *Journal of the American Statistical Association*, 91:1343–1370, 1996.

[28] K.S. Hall, A.O. Ogunniyi, H.C. Hendrie, B.O. Osuntokun, S.L. Hui, B. Musick, C.S. Rodenberg, F.W. Unverzagt, O. Guerje, and O. Baiyewu. A cross-cultural community based study of dementias: methods and performance of survey instrument. *International journal of methods in psychiatric research.*, 6:129–142, 1996.

[29] K.S. Hall, H.C. Hendrie, D.D. Rodgers, C.S. Prince, N. Pillay, A. Blue, H. Brittain, J.A. Norton, J.N. Kaufert, P. Shelton, , B.O. Osuntokun, and B.D. Postl. Development of dementia screening interview in two distinct languages. *International journal of methods in psychiatric research.*, 3:1–28, 1993.

[30] R.A. Murden, T.D. McRae, S. Kaner, and M.E. Buchnam. Mini-mental state exam scores vary with education in blacks and whites. *Journal of the American Geriatrics Society.*, 39:149–155, 1991.