HHS Public Access

Author manuscript

Epidemiol Method. Author manuscript; available in PMC 2015 April 02.

Published in final edited form as:

Epidemiol Method. 2013 September 1; 2(1): 49–66. doi:10.1515/em-2013-0008.

Extended Matrix and Inverse Matrix Methods Utilizing Internal Validation Data When Both Disease and Exposure Status Are Misclassified

Li Tang,

Department of Biostatistics, St. Jude Children's Research Hospital, Memphis, TN 38105, USA

Robert H. Lyles,

Department of Biostatistics and Bioinformatics, Rollins School of Public Health of Emory University, Atlanta, GA 30322, USA

Ye Ye,

Intelligent Systems Program and RODS Laboratory, University of Pittsburgh, Pittsburgh, PA 15206, USA

Yungtai Lo, and

Department of Epidemiology and Population Health, Albert Einstein College of Medicine, Bronx, NY 10461, USA

Caroline C. King

Division of Reproductive Health, Centers for Disease Control and Prevention, Atlanta, GA 30341, USA

Li Tang: li.tang@stjude.org; Robert H. Lyles: rlyles@emory.edu; Ye Ye: yey5@pitt.edu; Yungtai Lo: yungtai.lo@einstein.yu.edu; Caroline C. King: zpg2@cdc.gov

Abstract

The problem of misclassification is common in epidemiological and clinical research. In some cases, misclassification may be incurred when measuring both exposure and outcome variables. It is well known that validity of analytic results (e.g. point and confidence interval estimates for odds ratios of interest) can be forfeited when no correction effort is made. Therefore, valid and accessible methods with which to deal with these issues remain in high demand. Here, we elucidate extensions of well-studied methods in order to facilitate misclassification adjustment when a binary outcome and binary exposure variable are both subject to misclassification. By formulating generalizations of assumptions underlying well-studied "matrix" and "inverse matrix" methods into the framework of maximum likelihood, our approach allows the flexible modeling of a richer set of misclassification mechanisms when adequate internal validation data are available. The value of our extensions and a strong case for the internal validation design are demonstrated by means of simulations and analysis of bacterial vaginosis and trichomoniasis data from the HIV Epidemiology Research Study.

Correspondence to: Li Tang, `li.tang@stjude.org`.

Keywords

inverse matrix method; likelihood; matrix method; misclassification

1 Introduction

In many epidemiologic and clinical studies, one aims to quantify the association between binary disease and exposure status, for instance, via odds ratios (ORs) based on 2 × 2 tables. A common practical problem is that misclassification may exist in one or both variables. The threats to the validity of analytic results that stem from misclassification have received considerable attention. For example, the "matrix method" discussed in epidemiological textbooks (Kleinbaum et al., 1982; Rothman and Greenland, 1998) provides variations on an intuitive correction identity due to Barron (1977) that is parameterized in terms of familiar sensitivity and specificity properties of surrogate measurements on disease and exposure status. Greenland (1988) discussed point estimation and derived variance estimators under differential and nondifferential exposure misclassification using the matrix method, under various validation sampling schemes. By instead parameterizing in terms of positive and negative predictive values, Marshall (1990) developed an alternative correction identity later designated as the "inverse matrix method" (Morrissey and Spiegelman, 1999). The original inverse matrix method is restricted to the situation when there is differential misclassification of one variable (disease or exposure status), in which case it has been shown that Marshall's closed-form internal validation data-based corrected OR estimator is in fact a maximum-likelihood estimator (MLE) (Lyles, 2002; Greenland, 2008). Efficiency studies comparing the matrix and inverse matrix methods when exposure is misclassified also appear in the literature (Morrissey and Spiegelman, 1999).

We recognize the practical need of developing intuitive methods for estimating ORs in 2 × 2 tables with a more general view of misclassification. In particular, Barron's (1977) matrix method is an identity that assumes nondifferential and independent misclassification of both variables and is directly applicable only as a sensitivity analysis tool. Greenland and Kleinbaum (1983) extended this identity to permit differential but independent misclassification of both Y and X, but did not delve into efficient analysis based on validation data. Greenland (1988), Marshall (1990), Morrissey and Spiegelman (1999), and Lyles (2002) facilitated efficient estimation of the crude OR via validation data, but all considered misclassification of only one variable (e.g. exposure). Holcroft et al. (1997) tackled a similar problem with the use of a three-stage validation design, by proposing a class of semiparametric estimators.

Here, we seek to further extend the focus within the 2 × 2 table setting in a way that allows full generalization of the assumed misclassification process, and as a result subsumes the preceding treatments as special cases. This extension is driven by the practicalities of study design and analysis, as we focus on flexible modeling to account for complex misclassification via a rich internal validation sample when both binary variables are subject to errors in measurement. Rather than solely a theoretical exercise, it is directly motivated by real data for which we demonstrate that only this most general misclassification model is adequate.

In Section 2, we provide a maximum-likelihood (ML) framework that can be viewed as a practical facilitation of generalized versions of the matrix and inverse matrix methods. To our knowledge, it constitutes the first generalization of the matrix method identity to account for both dependent and differential misclassification and the first generalization of the inverse matrix identity to account for misclassification of both X and Y. We draw comparisons across methods and make suggestions for analyzing data in practice, heavily emphasizing the advantages of internal validation subsampling. This strategy, when feasible, facilitates efficient estimation of corrected ORs while avoiding serious biases that can occur when the assumed misclassification model is too simplistic. In addition, we suggest a model selection procedure that is readily implemented in standard statistical software. While our primary focus is on the point estimation of ORs in cross-sectional studies, we also briefly address the applicability of the methods to case–control studies. In Section 3, we introduce our motivating example, based on assessments of bacterial vaginosis (BV) and trichomoniasis (TRICH) in the HIV Epidemiology Research study (HERS). This example clearly illustrates how serious misinterpretation of the data can result when overly simplified misclassification models are assumed and highlights the benefits of the proposed approach. In Section 4, we present simulation studies to demonstrate the overall performance of the ML methodology in the context of cross-sectional studies.

2 Methods

2.1 Notation and terminology

2.1.1 Differential and dependent misclassification—Consider a 2 × 2 table in which one measures an error-prone surrogate X^* in place of a true exposure X and an error-prone Y^* in place of a true response Y. We assume X, X^*, Y, and Y^* are all binary variables. Now define $\pi_{xy} = \Pr(X = x, Y = y)$ and $\pi^*_{x^*y^*} = \Pr(X^* = x^*, Y^* = y^*)\ (x, y, x^*, y^* = 0, 1)$. The true OR of primary interest is given by $\pi_{11}\pi_{00}/\pi_{10}\pi_{01}$, while with misclassification in both variables, the naïve OR is $\pi^*_{11}\pi^*_{00} \Big/ \pi^*_{10}\pi^*_{01}$.

The observed data likelihood contribution for an observation with ($X^* = x^*$, $Y^* = y^*$) can be expressed as follows without losing generality:

$$\pi^*_{x^*y^*} = \sum_{x=0}^{1}\sum_{y=0}^{1} \Pr(Y^* = y^* | Y = y, X = x, X^* = x^*) Pr(X^* = x^* | X = x, Y = y)\pi_{xy}. \quad [1]$$

The first and second terms in eq. [1] represent the most general form of the likelihood expressed with a generalized version of the familiar misclassification parameters known as sensitivity (SE) and specificity (SP). Without additional constraints, we define $SE_{Yxx^*} = \Pr(Y^* = 1|Y = 1, X = x, X^* = x^*)$ and $SP_{Yxx^*} = \Pr(Y^* = 0|Y = 0, X = x, X^* = x^*)$. Note that misclassification parameters on Y depend on the joint distribution of (X, X^*), indicating the misclassification process in Y is differential but also depends on X, which is subject to misclassification too. This is potentially important, since it is far more common to assume independence of the misclassification processes (see Section 2.1.2). Similarly, denote $SE_{Xy} = \Pr(X^* = 1|X = 1, Y = y)$ and $SP_{Xy} = \Pr(X^* = 0|X = 0, Y = y)$, taking the typical form

associated with differential misclassification (Thomas et al., 1993). Terminology-wise, we view the general expression in eq. [1] as reflecting "differential and dependent misclassification".

Alternatively, one may choose to parameterize the observed data likelihood contribution in terms of positive and negative predictive values, that is,

$$\pi_{xy}=\sum_{x^*=0}^{1}\sum_{y^*=0}^{1}\Pr(Y=y|Y^*=y^*,X^*=x^*,X=x)\Pr(X=x|X^*=x^*,Y^*=y^*)\pi^*_{x^*y^*}, \quad [2]$$

where the first and second terms relate to predictive values of X and Y, defined as $PPV_{Yx}{}^* = \Pr(Y = 1|Y^* = 1, X = x, X^* = x^*)$, $NPV_{Yx}{}^* = \Pr(Y = 0|Y^* = 0, X = x, X^* = x^*)$, $PPV_{Xy}{}^* = \Pr(X = 1|X^* = 1, Y^* = y^*)$, and $NPV_{Xy}{}^* = \Pr(X = 0|X^* = 0, Y^* = y^*)$. In contrast to the parameterization using SE and SP, note that the predictive values of X depend on the potentially mismeasured response. Again, predictive values of Y depend on the joint distribution of (X, X^*), implying the dependence of misclassification of Y on the other misclassified variable. When only X is subject to misclassification, eq. [2] can be rewritten as $\pi_{xy}=\sum_{x^*=0}^{1}\Pr(X=x|X^*=x^*,Y=y)\Pr(X^*=x^*,Y=y)$. This reflects Marshall's (1990) original proposal, which we refer to as the "inverse matrix method".

2.1.2 Differential and independent misclassification—Assuming independent misclassification implies that $\Pr(Y^* = y^*, X^* = x^*|Y = y, X = x) = \Pr(Y^* = y^*|Y = y, X = x)\Pr(X^* = x^*|X = x, Y = y)$. In other words, X^* and Y^* are conditionally independent given (X, Y). However, it should be noted that the reverse may not be true. This corresponds to reducing eq. [1] to the following form:

$$\pi^*_{x^*y^*}=\sum_{x=0}^{1}\sum_{y=0}^{1}\Pr(Y^*=y^*|Y=y,X=x)\Pr(X^*=x^*|X=x,Y=y)\pi_{xy}, \quad [3]$$

where misclassification on Y only depends on true exposure X characterized by parameters $SE_{Yx} = \Pr(Y^* = 1|Y = 1, X = x)$ and $SP_{Yx} = \Pr(Y^* = 0|Y = 0, X = x)$. The misclassification model for X stays the same as in Section 2.1.1.

2.1.3 Nondifferential and independent misclassification—When assuming nondifferential and independent misclassification, we define $SE_X = \Pr(X^* = 1|X = 1)$, $SP_X = \Pr(X^* = 0|X = 0)$, $SE_Y = \Pr(Y^* = 1|Y = 1)$, and $SP_Y = \Pr(Y^* = 0|Y = 0)$. We can then rewrite the observed data likelihood contribution as:

$$\pi^*_{x^*y^*}=\sum_{x=0}^{1}\sum_{y=0}^{1}\Pr(Y^*=y^*|Y=y)\Pr(X^*=x^*|X=x)\pi_{xy}. \quad [4]$$

This corresponds to the setting originally studied by Barron (1977).

2.1.4 Other combinations—Sections 2.1.1–2.1.3 outline three misclassification mechanisms. However, other possibilities exist; for example, Y could be differentially but X nondifferentially misclassified. While we confine our main attention to the three situations

described above, the proposed methodology accommodates such variations without difficulty assuming adequate internal validation sampling.

2.2 ML approach

In general, the main study likelihood piece based on observed data pairs (Y_m^*, X_m^*) (m = 1, …, M) can be expressed as:

$$L_{\text{main}}=\prod_{m=1}^{M}\pi_{11}^{*(y_m^*x_m^*)}\pi_{01}^{*((1-x_m^*)y_m^*)}\pi_{10}^{*(x_m^*(1-y_m^*))}\pi_{00}^{*((1-x_m^*)(1-y_m^*))}, \quad [5]$$

where the π^*s take appropriate forms corresponding to different assumptions on the misclassification process as described in Section 2.1 and m denotes for the main study sample. For instance, if parameterizing in terms of SE/SP and allowing differential and dependent misclassification, we have $\pi_{11}^*=\text{SE}_{Y11}\pi_{11}\text{SE}_{X1}+\text{SE}_{Y01}\pi_{01}(1-\text{SP}_{X1})+(1-\text{SP}_{Y11})\pi_{10}\text{SE}_{X0}+(1-\text{SP}_{Y01})\pi_{00}(1-\text{SP}_{X0})$. In contrast, if independence is assumed while preserving differentiality on both variables, $\pi_{11}^*=\text{SE}_{Y1}\pi_{11}\text{SE}_{X1}+\text{SE}_{Y0}\pi_{01}(1-\text{SP}_{X1})+(1-\text{SP}_{Y1})\pi_{10}\text{SE}_{X0}+(1-\text{SP}_{Y0})\pi_{00}(1-\text{SP}_{X0})$. Under the most simplified setting (e.g. Barron, 1977), the simultaneous assumptions of independent and nondifferential misclassification imply that $\pi_{11}^*=\text{SE}_Y\pi_{11}\text{SE}_X+\text{SE}_Y\pi_{01}(1-\text{SP}_X)+(1-\text{SP}_Y)\pi_{10}\text{SE}_X+(1-\text{SP}_Y)\pi_{00}(1-\text{SP}_X)$. The other π^*s are derived similarly under each scenario (Tang, 2012). Note that the “main study only” likelihood in eq. [5] is directly applicable solely for sensitivity analysis. We emphasize extensions to accommodate a main/internal validation design in Section 2.5.

2.3 Generalized matrix method

We generalize the concept of the matrix method and its extensions (Kleinbaum et al., 1982; Greenland and Kleinbaum, 1983) by flexibly incorporating the full range of possible misclassification models. In general, one is able to relate surrogate and true cell probabilities via the equality $\boldsymbol{\Pi}^* = \boldsymbol{A}\boldsymbol{\Pi}$, where $\boldsymbol{\Pi}=(\pi_{11}\ \pi_{01}\ \pi_{10}\ \pi_{00})'$, $\Pi^*=(\ \pi_{11}^*\ \ \pi_{01}^*\ \ \pi_{10}^*\ \ \pi_{00}^*\ )'$ and the definition of $\boldsymbol{A}$ varies according to the assumptions made. For differential and dependent misclassification, we derive $\boldsymbol{A}$ in its most general form as follows:

$$\boldsymbol{A}=\begin{bmatrix} \text{SE}_{Y11}\text{SE}_{X1} & \text{SE}_{Y01}(1-\text{SP}_{X1}) & (1-\text{SP}_{Y11})\text{SE}_{X0} & (1-\text{SP}_{Y01})(1-\text{SP}_{X0}) \\ \text{SE}_{Y10}(1-\text{SE}_{X1}) & \text{SE}_{Y00}\text{SP}_{X1} & (1-\text{SP}_{Y10})(1-\text{SE}_{X0}) & (1-\text{SP}_{Y00})\text{SP}_{X0} \\ (1-\text{SE}_{Y11})\text{SE}_{X1} & (1-\text{SE}_{Y01})(1-\text{SP}_{X1}) & \text{SP}_{Y11}\text{SE}_{X0} & \text{SP}_{Y01}(1-\text{SP}_{X0}) \\ (1-\text{SE}_{Y10})(1-\text{SE}_{X1}) & (1-\text{SE}_{Y00})\text{SP}_{X1} & \text{SP}_{Y10}(1-\text{SE}_{X0}) & \text{SP}_{Y00}\text{SP}_{X0} \end{bmatrix}$$

Under other assumptions, the matrix $\boldsymbol{A}$ can be derived as in Appendix 1. The matrix method identity relies upon inversion of the matrix $\boldsymbol{A}$ in order to obtain the vector $\boldsymbol{\Pi}=\boldsymbol{A}^{-1}\boldsymbol{\Pi}^*$.

2.4 Generalized inverse matrix method

The inverse matrix identity directly expresses true cell probabilities as sums of products of surrogate cell probabilities and predictive values. Here, we extend the proposal of Marshall (1990) to a general context with both variables misclassified in a 2 × 2 table. For example, under dependent and differential misclassification, the law of total probability dictates that

$\pi_{11}=\mathrm{PPV}_{Y11}\pi^*_{11}\mathrm{PPV}_{X1}+(1-\mathrm{NPV}_{Y11})\pi^*_{10}\mathrm{PPV}_{X0}+\mathrm{PPV}_{Y10}\pi^*_{01}(1-\mathrm{NPV}_{X1})+(1-\mathrm{NPV}_{Y10})\pi^*_{00}(1-\mathrm{NPV}_{X0})$. Packaging linear equations into matrices, the form of the generalized inverse matrix method is as given in Marshall's original proposal: $\boldsymbol{\Pi}=\boldsymbol{B}\boldsymbol{\Pi}^*$. However, in our approach, the matrix $\boldsymbol{B}$ takes a more complicated form to accommodate a general misclassification mechanism for both the *X* and the *Y* variables:

$$\boldsymbol{B}=\begin{bmatrix} \mathrm{PPV}_{Y11}\mathrm{PPV}_{X1} & \mathrm{PPV}_{Y10}(1-\mathrm{NPV}_{X1}) & (1-\mathrm{NPV}_{Y11})\mathrm{PPV}_{X0} & (1-\mathrm{NPV}_{Y10})(1-\mathrm{NPV}_{X0}) \\ \mathrm{PPV}_{Y01}(1-\mathrm{PPV}_{X1}) & \mathrm{PPV}_{Y00}\mathrm{NPV}_{X1} & (1-\mathrm{NPV}_{Y01})(1-\mathrm{PPV}_{X0}) & (1-\mathrm{NPV}_{Y00})\mathrm{NPV}_{X0} \\ (1-\mathrm{PPV}_{Y11})\mathrm{PPV}_{X1} & (1-\mathrm{PPV}_{Y10})(1-\mathrm{NPV}_{X1}) & \mathrm{NPV}_{Y11}\mathrm{PPV}_{X0} & \mathrm{NPV}_{Y10}(1-\mathrm{NPV}_{X0}) \\ (1-\mathrm{PPV}_{Y01})(1-\mathrm{PPV}_{X1}) & (1-\mathrm{PPV}_{Y00})\mathrm{NPV}_{X1} & \mathrm{NPV}_{Y01}(1-\mathrm{PPV}_{X0}) & \mathrm{NPV}_{Y00}\mathrm{NPV}_{X0} \end{bmatrix}$$

In contrast to the generalized matrix method, there is no matrix inversion involved in computing the corrected OR through the generalized inverse matrix method. In principle, this could confer a numerical advantage in practice, although again direct use of the identity is generally restricted to the setting of sensitivity analysis.

2.5 Estimation via internal validation sampling

The estimate of the corrected OR is $\hat{\mathrm{OR}}=\hat{\pi_{11}}\hat{\pi_{00}}\big/\hat{\pi_{10}}\hat{\pi_{01}}$. For all of the approaches presented above, estimation of misclassification probabilities is crucial in practice. When possible, we recommend the use of an internal validation subsample randomly selected from one's current study, for which both true binary variables are measured via gold-standard methods along with the error-prone methods used in the main study. The primary appeal of adopting internal (as opposed to external) validation sampling is the avoidance of the necessity to assume "transportability" of misclassification probabilities (Begg, 1987; Carroll et al., 2006) and the accommodation of more general misclassification mechanisms.

When allowing full generality, that is, dependent and differential misclassification, it can be shown that a full likelihood approach based on the proposed main/internal validation design is equivalent regardless of whether parameterized based on predictive values or SE/SP probabilities (Tang, 2012). There are in total 16 types of validation set records, if validations on *X* and *Y* are measured simultaneously for each subject in the subsample. Table 1 shows the likelihood contributions for each validation record type based on both parameterizations. In contrast, the main study likelihood based on (X^*, Y^*) records is given explicitly in eq. [5], that is,

$$L_{\mathrm{main}}=\prod_{m=1}^{M}\pi_{11}^{*(y^*_m x^*_m)}\pi_{01}^{*((1-x^*_m)y^*_m)}\pi_{10}^{*(x^*_m(1-y^*_m))}\pi_{00}^{*((1-x^*_m)(1-y^*_m))}.$$

If parameterizing in terms of SE and SP values, all the π^*s are further expanded (see Section 2.2).

The internal validation subsample likelihood is given by

$$L_{\text{val}}=\prod_{p=1}^{16} L_{v_p}^{n_{v_p}},$$

where L_{vp} is the likelihood term corresponding to observation type p in Table 1, while n_{vp} is the total number of observations of the pth type (p = 1, 2, … 16). Note that the total validation study sample size is $n_v=\sum_{p=1}^{16} n_{v_p}$. The overall likelihood to be maximized is based on a total of $M + n_v$ subjects and is proportional to the product of the main and validation study components, i.e. $L_{\text{main}} \times L_{\text{val}}$.

There are no closed-form solutions for the MLEs based on the overall likelihood written in terms of SE and SP. Interestingly, however, closed forms exist for the predictive value parameterization in the most general case. For example, one can readily verify that

$$\hat{\pi}_{11}^{*}=\frac{\sum_{i=1}^{M+n_v} I_{X_i^*=1,Y_i^*=1}}{M+n_v} \text{ and } \hat{\text{PPV}}_{Y11}=\frac{\sum_{i=1}^{n_v} I_{\text{val}=1,y_i^*=1,y_i=1,x_i=1,x_i^*=1}}{\sum_{i=1}^{n_v} I_{\text{val}=1,y_i^*=1,x_i=1,x_i^*=1}},$$

where the I notation represents an indicator that the conditions described in the subscript are met (Tang, 2012). The MLEs for the πs can then be estimated from the $\hat{\pi}^*$s, $\hat{\text{PPV}}_{\text{S}}$, and $\hat{\text{NPV}}_{\text{S}}$ by direct use of the generalized inverse matrix identity of Section 2.4. Because the two parameterizations are equivalent under the circumstance of dependent and differential misclassification, we may also obtain closed-form MLEs for the $\hat{\text{SE}}$ and $\hat{\text{SP}}$ parameters as functions of the $\hat{\text{PPV}}_{\text{S}}$ and $\hat{\text{NPV}}_{\text{S}}$ in that setting. For example,

$$\hat{\text{SE}}_{Y11}=\frac{\hat{\text{PPV}}_{Y11}\hat{\text{PPV}}_{X1}\hat{x}_{11}^{*}}{\hat{\text{PPV}}_{Y11}\hat{\text{PPV}}_{X1}\hat{\pi}_{11}^{*}+(1-\hat{\text{NPV}}_{Y11})\hat{\text{PPV}}_{X0}\hat{\pi}_{10}^{*}}.$$

The remaining closed-form MLEs are displayed in Appendix 2.

When the misclassification process is not fully general (e.g. assuming independent misclassification and/or nondifferential misclassification of either variable), the equivalence between the likelihoods based on the SE/SP and predictive value parameterizations no longer holds. In such cases, it appears that there are no simple closed forms for likelihood-based $\hat{\text{SE}}$s, $\hat{\text{SP}}$s, and $\hat{\pi}$s. If one supplies the generalized matrix method with data-driven SE and SP estimates that are not MLEs, the corrected $\hat{\text{OR}}$ will not be fully efficient. These conclusions are consistent with previous findings in a simpler context, with misclassification of only one variable (Lyles, 2002).

In general, we recommend the use of the ML approach for optimal efficiency and the ease of numerically computing standard errors. Optimizing the full main/internal validation likelihood under either parameterization path is readily achieved by taking advantage of numerical procedures in standard statistical software. As such, we view the matrix and

inverse matrix constructs more as instructive identities than as practical analysis tools, unless they are to be used solely for sensitivity analyses. Straightforward multivariate delta-method calculations allow computing the approximate standard error of the corrected $\log(\hat{\mathrm{OR}})$ based on ML, after obtaining the $\hat{\pi}$s and the corresponding numerically-derived Hessian. SAS NLMIXED (SAS Institute, Inc., 2008) programs for accomplishing these tasks are readily available from the first author.

A natural question one might ask is whether measuring (X^*, Y^*) on every subject in addition to (X, Y) yields a different or improved estimate of the true OR characterizing the (X, Y) association. In fact, if (X, Y, X^*, Y^*) is available on all participants, the available information for estimating the OR is equivalent to that contained in the (X, Y) data alone. The overall likelihood then reduces to L_{val}. In Appendix 3, we show that maximizing the reduced form of the overall likelihood (L_{val} only) under the most general misclassification model in this situation leads to exactly the same MLEs of the πs as those obtained from analyses ignoring (X^*, Y^*). A similar argument can be readily derived under other types of misclassification models. This finding unsurprisingly suggests that knowing surrogates when gold-standard measures are available on the whole sample does not offer additional value in the estimation of the primary effect (e.g. OR) of interest, which further implies that if gold-standard measures are comparatively affordable compared to surrogates, it is more efficient to evaluate via gold standards only.

2.6 Notes on case–control studies

While our focus has been on cross-sectional sampling, the case–control sampling scheme is also worthy of discussion. Here, we consider "case–control" studies as those where case oversampling is conducted based on the error-prone responses. In other words, observations with $Y^* = 1$ ("cases") are sampled with a greater probability than those with $Y^* = 0$ ("controls"). Prior work (Greenland and Kleinbaum, 1983) has noted that supplying the population misclassification probabilities to the correction methods will yield invalid estimates; however, with nondifferential misclassification, the validity of the analytic results could be restored by introducing the sampling fraction of cases and controls into the correction. It was also noted in Lyles et al. (2011) that the main/internal validation design is favorable for handling such oversampling under nondifferential misclassification, because it automatically yields estimates of the "operating" misclassification probabilities. Similar findings are observed in the current setting. With oversampling of "cases" ($Y^* = 1$), the method described in the previous sections yields valid estimation of the OR, as long as misclassification of Y is nondifferential. When the nondifferential misclassification assumption is not met, however, the validity of the estimated OR based on the main/internal validation design does not hold under "case" oversampling. More details can be found in Tang (2012).

2.7 Model selection

When correcting the estimate of the OR, we would ideally choose the misclassification mechanism that generated the observed data. Here, we provide a straightforward model selection procedure to guide practitioners. For ease of discussion, denote the dependent and differential misclassification model as "Model 1", followed by "Model 2" (the independent

and differential misclassification model in Section 2.1.2) and “Model 3” (the completely nondifferential model in Section 2.1.3). Model 1 reflects a fully general misclassification mechanism, while Model 2 can be regarded as a generalization of Marshall’s (1990) framework to the situation when both X and Y are misclassified and Model 3 is a representation of Barron’s (1977) setting.

Define AIC_q = the value of the Akaike Information Criterion (AIC) (Akaike, 1974) upon fitting Model q (q = 1, 2, 3). In practice, we recommend selecting the model that yields the smallest value of AIC, as that criterion is well known to balance between the number of necessary parameters included and the quality of model fit. One may then simply report the results corresponding to the selected model. Although a more accurate standard error for the resulting estimated log(OR) might presumably be obtained via resampling, our empirical studies suggest that it is suitably reliable and computationally efficient to report the standard error from the selected model (see Section 4). We apply this AIC-based approach to real data in the following section, and a program utilizing the SAS NLMIXED procedure to implement the model selection method is available from the first author. For additional comments regarding selection of the misclassification model, see Section 5.

3 Example

Our motivating example comes from the HERS. This is a multi-center prospective cohort study with a total of 1,310 women enrolled in four U.S. cities from 1993 to 1995 (Smith et al., 1997). Among them, 871 women were HIV-infected, and 439 were not infected but at risk. During each semi-annual visit, a wealth of subject-specific information was collected. The question of interest is to assess the association between two binary variables: BV status and TRICH status. BV was measured by two different clinical methods: the clinically-based (CLIN) and the laboratory-based (LAB) methods. CLIN is a less accurate method that diagnoses BV by evaluating multiple clinical criteria based on a modified Amsel’s criteria (Amsel et al., 1983), while LAB relies on a more sophisticated Gram-staining technique (Nugent et al., 1991). The LAB method is more expensive and serves here as an arguable gold standard, while the CLIN method is more cost-efficient and accessible. The presence of TRICH was evaluated by a clinical wet mount technique characterized by low sensitivity (Thomason et al., 1988), along with a gold-standard culture method. For both BV and TRICH measurements, gold-standard and error-prone diagnoses are widely available for HERS participants at Visit 4 and beyond. This feature of the HERS makes for an excellent illustrative example of internal validation data-based methodology.

We consider 916 patients with complete observations on both error-prone and gold-standard diagnoses of BV and TRICH at the fourth HERS visit. We selected Visit 4, because a previous examination uncovered a complex misclassification process underlying the assessment of BV status at that visit (Lyles et al., 2011). The prevalence of BV via the LAB technique in the sample was 18.2%, while due to misclassifying some diagnoses the naïve CLIN prevalence was only 7.5%. Compared to the LAB BV diagnosis, estimates suggest that CLIN BV conferred a crude SE around 37% and a crude SP of about 99%. The prevalence of TRICH in our sample was 40.2% when assessed by culture testing. In

contrast, when evaluated by wet mount, the prevalence was only 24.5%, with an estimated crude SE of 51.9% and SP of 94.0%.

Table 2 summarizes the results based on using gold-standard measurements only, error-prone diagnoses only, and fitting correction models via the proposed main/internal validation design under various misclassification mechanisms. Note that the naïve result characterizing the association between CLIN BV and wet mount-based TRICH inflated the estimated OR by nearly 50% relative to the LAB and culture-based analyses. For main/internal validation analysis based on Models 1–3, we utilized a random subsample selecting ¼ of the total sample size as the internal validation set. A summary of the data comprising the resulting main and internal validation samples is presented in Table 7 (Appendix 4). The corrected $\widehat{\mathrm{OR}}$ is close to the gold-standard (LAB and culture-based) result, though with expected efficiency loss, when dependent and differential misclassification is allowed (Model 1). If differential but independent misclassification (Model 2) is assumed, the corrected $\widehat{\mathrm{OR}}$ appears slightly biased away from the null. When a nondifferential misclassification model is adopted, the corrected $\widehat{\mathrm{OR}}$ is similar to that obtained via the naïve result.

With the proposed model selection approach (Section 2.7), Model 1 is chosen with the smallest AIC value among the three candidate models. Therefore, we retain the fully general Model 1 as the final model, suggesting that the HERS data require one to account for dependent misclassification that is differential with respect to both X and Y. The results indicate that TRICH is positively associated with BV among the HERS population at Visit 4, and our corrected analysis based on Model 1 agrees with the gold-standard analysis extremely well.

As discussed in Section 2.5, when utilizing both the gold-standard and surrogate measures of BV and TRICH for all 916 subjects in order to specify the corresponding full likelihood L_{val}, we obtained the identical log(OR) estimate and standard error as when performing the "gold-standard" analysis in Table 2. Therefore, this result is omitted from the table.

4 Simulation studies

4.1 Study I: mimicking real-data example

Our first simulation experiment evaluates the performance of the proposed methods under conditions mimicking the HERS example (Section 3). Cell counts were simulated from a multinomial distribution with cell probabilities of (π_{11} = 0.1146, π_{10} = 0.2871, π_{01} = 0.0677, π_{00} = 0.5306), and main and internal validation sample sizes (n_m = 687, n_v = 219) similar to those observed in the HERS example. Error-prone response Y^* and exposure X^* were generated with misclassification probabilities estimated from the HERS sample based on the fit of Model 1 (data available in Table 7), where the misclassification process was assumed dependent and differential. For each of 500 simulated datasets, we conducted naïve analysis associating Y^* with X^*, true analysis with Y and X, and main/internal validation analyses via Models 1–3.

Table 3 summarizes the results. The naïve analysis yields a result biased away from the null. Model 1 produces the corrected OR estimate closest to the gold-standard OR, with tolerable sacrifice in efficiency. The 95% CI coverage under Model 1 is also excellent. When reducing Model 1 to other simpler versions by assuming independent or nondifferential misclassification, the results are biased, reflecting the fact that the reduced models are not consistent with the data generation process. Note that with the simplest model assuming nondifferential misclassification of both variables (Model 3), the corrected result is similar to the naïve result (in fact, arguably worse). This strongly highlights the importance of internal validation data to permit flexibility in the selected misclassification model.

The corrected results using the generalized matrix methods discussed in Section 2.1.1 agree well with the MLEs, when ML estimates of misclassification probabilities are supplied. However, when simpler crude estimates obtained from the validation subsample are inserted into the generalized matrix method, results are not satisfying, even producing negative estimates of probabilities in some cases (Tang, 2012; results not shown). Thus, in practice, we favor the proposed main/internal validation study-based full ML approach in the interest of obtaining both valid and efficient results.

4.2 Study II: performance of model selection

The results in Section 4.1 suggest the importance of misclassification model selection to ensure the model is specified correctly (or, at least, generally enough). Extensive simulations were performed to evaluate the performance of the proposed AIC-based model selection strategy (Tables 4–6), when the underlying association was negative (Table 4), or moderate positive (Table 5), or strong positive (Table 6). Under various settings, the model was chosen correctly most of the time. For example, under setting 4, the true underlying model from which data were generated was Model 3. Unsurprisingly, the more general Models 1 and 2 yield valid results. However, with the proposed model selection strategy, Model 3 is correctly picked 88.0% of the time, yielding a slight improvement in efficiency relative to Model 1. In contrast, under setting 6, Model 1 is the underlying model; thus, estimates from Models 2 and 3 are not valid. By correctly selecting Model 1, 94.8% of the time, however, the model selection strategy maintained overall validity and achieved satisfactory 95% CI coverage.

The simulation results in Tables 4–6 suggest that AIC is a highly effective criterion for selecting among the alternative misclassification models. The key concern, however, is maintenance of validity in the OR estimate. Since the true misclassification model is unknown, only Model 1 ensures such validity in theory. Thus, whenever the internal validation subsample is of adequate size to support its fit, Model 1 must be viewed as the safest choice. Another argument in favor of Model 1 is the fact that, at least under the simulation conditions examined here, it produced a log(OR) estimate with very similar mean and variance properties to those characterizing the MLEs under simpler true underlying misclassification models.

5 Discussion

We have considered the classic problem of analyzing 2 × 2 tables, when both binary variables are subject to misclassification. Our main contributions are twofold. First, we have expanded the well-studied matrix (Barron, 1977) and inverse matrix (Marshall, 1990) identities to a more general context than ever before. Specifically, the results given in Sections 2.3 and 2.4 extend both identities to a fully general scenario with dependent and differential misclassification of two binary variables and could serve to update epidemiological methodology texts with regard to this topic. Secondly, we place heavy emphasis on specifying likelihood functions corresponding to main/internal validation designs under potentially complex misclassification mechanisms involving two binary variables. To our knowledge, this effort provides the first fully articulated framework to accomplish a joint main/internal validation study-based ML analysis allowing for dependent and differential misclassification of both variables. By parameterizing in terms of positive and negative predictive values, we have derived closed-form MLEs for the true cell probabilities based on this fully general misclassification model. The ML analysis requires numerical optimization under more restrictive nested misclassification models, but easily implemented programs designed to fit Models 1–3 (Section 2.7) using SAS NLMIXED are available from the first author by request.

In the context studied here, the ability to apply a misclassification model that is sufficiently general can be critical, if one hopes to obtain a valid estimate of association. Our motivating example involving BV and TRICH assessments from the HERS illustrates this point extremely well, as we find evidence suggesting bias in all estimates of the OR except the one based on the fully general dependent and differential misclassification model introduced in this article. When misclassification of either variable is differential, the naïve log(OR) estimator can be biased in either direction. Moreover, the HERS example demonstrates that a corrected estimate based on an incorrect nondifferential error assumption for either variable can be potentially worse than the naïve estimate. For this reason, we urge practitioners not to simply assume nondifferential misclassification of either variable, unless that assumption is supported by the data or there is no other resource.

It should be noted that familiar matrix and inverse matrix methods as applied in practice are only equivalent to special cases of the proposed likelihood-based approach, when MLEs of misclassification rates are supplied into the generalized matrix identities. Otherwise, estimators based on application of the matrix and inverse matrix methods are not fully efficient. For this reason, we favor the approach advocated here in which the full main/internal validation study likelihood is utilized. If one is also interested in obtaining a confidence interval for the OR, numerical optimization of the likelihood function greatly reduces the complexity of delta-method-based calculations for computing standard errors to accompany the adjusted log(OR) estimate (Tang, 2012; details and program available from first author).

We have proposed a straightforward model selection procedure for practitioners who not only seek to obtain a valid analytic result but also pursue a more precise result that may be achievable via a correct reduced misclassification model. It has been demonstrated that the

proposed model selection procedure works stably and permits the choice of simpler models when the deviation of the estimated OR is acceptable relative to the general model. However, since the saturated model allowing dependent and differential misclassification is always valid and appeared to sacrifice little efficiency in our simulations given an adequate validation sample, it may often be prudent to avoid model selection and simply settle upon the saturated misclassification model.

Our findings suggest that when designing large-scale epidemiologic studies for which standard outcome (Y) and exposure (X) assessments are error-prone, it is valuable to invest in collecting an internal validation subsample with gold-standard measurements applied to both Y and X. This allows one to evaluate and adjust for differential and/or dependent misclassification if it could be an issue. When gold standards are not available, however, one should consider sensitivity analyses to explore the potential effects of misclassification (Lash and Fink, 2003; Fox et al., 2005; Lyles and Lin, 2010). In our context, a series of pre-specified misclassification rates could be supplied into matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ of the generalized matrix and inverse matrix methods in Sections 2.3 and 2.4, respectively, to assess their impact on the estimated OR. We caution, however, that such sensitivity analyses may generally be invalid under case oversampling (e.g. Greenland and Kleinbaum, 1983).

We are currently investigating natural extensions of the current work to the multivariable regression and longitudinal settings, with internal validation subsampling to facilitate misclassification adjustments. Future work could involve specific consideration of cost-efficient internal validation designs when both X and Y are misclassified, as in practice the costs associated with validating X or Y may be different. As an extension of prior work, it could be of interest to consider the allocation of validated observations cleverly into different types, to ensure the control of cost while still maintaining analytic validity. In some cases, formal considerations of this question may reveal the most cost-efficient approach to be the one in which the gold-standard approach is applied to all experimental units (Spiegelman and Gray, 1991; Lyles et al., 2005). A sample simulation program evaluating analytic validity with various validation sample sizes and pre-specified parameters is available from the author upon request, which offers a practical guide for study planning. Also, investigators may sometimes be more interested in validating a particular subpopulation, for example, those with a disease than those without, leading to nonrandom validation sampling. There could also be interest in extending the methods studied here to settings in which one or both gold-standard methods are imperfect, or “alloyed” (Wacholder et al., 1993; Brenner, 1996).

References

Akaike H. A new look at the statistical model identification. IEEE Transactions on Automatic Control. 1974; 19:716–723.

Amsel R, Totten PA, Spiegel CA, Chen KC, Eschenbach D, Holmes KK. Nonspecific vaginitis: diagnostic criteria and microbial and epidemiologic associations. American Journal of Medicine. 1983; 74:14–22. [PubMed: 6600371]

Barron BA. The effects of misclassification on the estimation of relative risk. Biometrics. 1977; 33:414–418. [PubMed: 884199]

Begg CB. Biases in the assessment of diagnostic tests. Statistics in Medicine. 1987; 6:411–423. [PubMed: 3114858]

Brenner H. Correcting for exposure misclassification using an alloyed gold standard. Epidemiology. 1996; 7:406–410. [PubMed: 8793367]

Carroll, RJ.; Ruppert, D.; Stefanski, LA. Measurement Error in Nonlinear Models. 2. London: Chapman and Hall; 2006.

Fox MP, Lash TL, Greenland S. A method to automate probabilistic sensitivity analyses of misclassified binary variables. International Journal of Epidemiology. 2005; 34:1370–1376. [PubMed: 16172102]

Greenland S. Variance estimation for epidemiologic effect estimates under misclassification. Statistics in Medicine. 1988; 7:745–757. [PubMed: 3043623]

Greenland S. Maximum-likelihood and closed-form estimators of epidemiologic measures under misclassification. Journal of Statistical Planning and Inference. 2008; 138:528–538.

Greenland S, Kleinbaum D. Correcting for misclassification in two-way tables and matched-pair studies. International Journal of Epidemiology. 1983; 12:93–97. [PubMed: 6840961]

Holcroft CA, Rotnitzky A, Robins JM. Efficient estimation of regression parameters from multistage studies with validation of outcomes and covariates. Journal of Statistical Planning and Inference. 1997; 65:349–374.

SAS Institute Inc. SAS/STAT® 9.2 User's Guide. Cary, NC: SAS Institute Inc; 2008.

Kleinbaum, D.; Kupper, L.; Morgenstern, H. Epidemiologic Research: Principles and Quantitative Methods. Belmont, CA: Lifetime Learning; 1982.

Lash TL, Fink AK. Semi-automated sensitivity analysis to assess systematic errors in observational data. Epidemiology. 2003; 14:451–458. [PubMed: 12843771]

Lyles RH. A note on estimating crude odds ratios in case–control studies with differentially misclassified exposure. Biometrics. 2002; 58:1034–1037. [PubMed: 12495160]

Lyles RH, Lin J. Sensitivity analysis for misclassification in logistic regression via likelihood methods and predictive value weighting. Statistics in Medicine. 2010; 29:2297–2309. [PubMed: 20552681]

Lyles RH, Tang L, Superak HM, King CC, Celantano D, Lo Y, Sobel J. Validation data-based adjustments for outcome misclassification in logistic regression: an illustration. Epidemiology. 2011; 22:589–597. [PubMed: 21487295]

Lyles RH, Williamson JM, Lin HM, Heilig CM. Extending McNemar's test: estimation and inference when paired binary outcome data are misclassified. Biometrics. 2005; 61:281–294.

Marshall RJ. Validation study methods for estimating proportions and odds ratios with misclassified data. Journal of Clinical Epidemiology. 1990; 43:941–947. [PubMed: 2213082]

Morrissey MJ, Spiegelman D. Matrix methods for estimating odds ratios with misclassified exposure data: extensions and comparisons. Biometrics. 1999; 55:338–344. [PubMed: 11318185]

Nugent RP, Krohn MA, Hillier SL. Reliability of diagnosing bacterial vaginosis is improved by a standardized method of gram stain interpretation. Journal of Clinical Microbiology. 1991; 29:297–301. [PubMed: 1706728]

Rothman, KJ.; Greenland, S. Modern Epidemiology. Philadelphia, PA: Lippincott-Raven; 1998.

Smith DK, Warren DL, Vlahov D, Schuman P, Stein MD, Greenberg BL. Design and baseline participant characteristics of the Human Immunodeficiency Virus Epidemiology Research (HER) Study: a prospective cohort study of human immunodeficiency virus infection in U.S. women. American Journal of Epidemiology. 1997; 146:459–469. [PubMed: 9290506]

Spiegelman D, Gray R. Cost-efficient study designs for binary response data with generalized Gaussian measurement error in the covariate. Biometrics. 1991; 47:851–870. [PubMed: 1789885]

Tang, L. PhD Dissertation. Atlanta, GA: Emory University; 2012. Analysis of Data with Complex Misclassification in Response or Predictor Variables by Incorporating Validation Subsampling.

Thomas D, Stram D, Dwyer J. Exposure measurement error: influence on exposure-disease. Relationships and methods of correction. Annual Review of Public Health. 1993; 14:69–93.

Thomason JL, Gelbart SM, Sobun JF, Schulien MB, Hamilton PR. Comparison of four methods to detect Trichomonas vaginalis. Journal Clinical Microbiology. 1988; 26:1869–1870.

Wacholder S, Armstrong B, Hartge P. Validation studies using an alloyed gold standard. American Journal of Epidemiology. 1993; 137:1251–1258. [PubMed: 8322765]

Appendix 1: matrix A for generalized matrix identity under various situations

Assuming differential misclassification with independence,

$$A=\begin{bmatrix} \mathrm{SE}_{Y1}\mathrm{SE}_{X1} & \mathrm{SE}_{Y0}(1-\mathrm{SP}_{X1}) & (1-\mathrm{SP}_{Y1})\mathrm{SE}_{X0} & (1-\mathrm{SP}_{Y0})(1-\mathrm{SP}_{X0}) \\ \mathrm{SE}_{Y1}(1-\mathrm{SE}_{X1}) & \mathrm{SE}_{Y0}\mathrm{SP}_{X1} & (1-\mathrm{SP}_{Y1})(1-\mathrm{SE}_{X0}) & (1-\mathrm{SP}_{Y0})\mathrm{SP}_{X0} \\ (1-\mathrm{SE}_{Y1})\mathrm{SE}_{X1} & (1-\mathrm{SE}_{Y0})(1-\mathrm{SP}_{X1}) & \mathrm{SP}_{Y1}\mathrm{SE}_{X0} & \mathrm{SP}_{Y0}(1-\mathrm{SP}_{X0}) \\ (1-\mathrm{SE}_{Y1})(1-\mathrm{SE}_{X1}) & (1-\mathrm{SE}_{Y0})\mathrm{SP}_{X1} & \mathrm{SP}_{Y1}(1-\mathrm{SE}_{X0}) & \mathrm{SP}_{Y0}\mathrm{SP}_{X0} \end{bmatrix}$$

which has the same form as defined by Greenland and Kleinbaum (1983). Under the circumstance of nondifferential and independent misclassification,

$$A=\begin{bmatrix} \mathrm{SE}_{Y}\mathrm{SE}_{X} & \mathrm{SE}_{Y}(1-\mathrm{SP}_{X}) & (1-\mathrm{SP}_{Y})\mathrm{SE}_{X} & (1-\mathrm{SP}_{Y})(1-\mathrm{SP}_{X}) \\ \mathrm{SE}_{Y}(1-\mathrm{SE}_{X}) & \mathrm{SE}_{Y}\mathrm{SP}_{X} & (1-\mathrm{SP}_{Y})(1-\mathrm{SE}_{X}) & (1-\mathrm{SP}_{Y})\mathrm{SP}_{X} \\ (1-\mathrm{SE}_{Y})\mathrm{SE}_{X} & (1-\mathrm{SE}_{Y})(1-\mathrm{SP}_{X}) & \mathrm{SP}_{Y}\mathrm{SE}_{X} & \mathrm{SP}_{Y}(1-\mathrm{SP}_{X}) \\ (1-\mathrm{SE}_{Y})(1-\mathrm{SE}_{X}) & (1-\mathrm{SE}_{Y})\mathrm{SP}_{X} & \mathrm{SP}_{Y}(1-\mathrm{SE}_{X}) & \mathrm{SP}_{Y}\mathrm{SP}_{X} \end{bmatrix}$$

and with some algebraic work, one can easily show that this equation is equivalent to that underlying Barron's original matrix method (Barron, 1977). With algebraic work, it can be shown that $\boldsymbol{A}$ is invertible if and only if $\mathrm{SE}_X + \mathrm{SP}_X - 1 > 0$ and $\mathrm{SE}_Y + \mathrm{SP}_Y - 1 > 0$. Under usual circumstances with reasonable error-prone assessments, one can reasonably expect these two inequalities to hold. The generalized matrix method is then derived immediately as $\boldsymbol{\Pi} = \boldsymbol{A}^{-1}\boldsymbol{\Pi}^{*}$.

Appendix 2: closed-form ML estimators for SE and SP parameters

$$\hat{\mathrm{SE}}_{Y11}=\frac{\hat{\mathrm{PPV}}_{Y11}\hat{\mathrm{PPV}}_{X1}\hat{\pi}^*_{11}}{\hat{\mathrm{PPV}}_{Y11}\hat{\mathrm{PPV}}_{X1}\hat{\pi}^*_{11}+(1-\hat{\mathrm{NPV}}_{Y11})\hat{\mathrm{PPV}}_{X0}\hat{\pi}^*_{10}}$$

$$\hat{\mathrm{SE}}_{Y10}=\frac{\hat{\mathrm{PPV}}_{Y10}(1-\hat{\mathrm{NPV}}_{X1})\hat{\pi}^*_{11}}{\hat{\mathrm{PPV}}_{Y10}(1-\hat{\mathrm{NPV}}_{X1})\hat{\pi}^*_{01}+(1-\hat{\mathrm{NPV}}_{Y10})(1-\hat{\mathrm{NPV}}_{X0})\hat{\pi}^*_{00}}$$

$$\hat{\mathrm{SE}}_{Y01}=\frac{\hat{\mathrm{PPV}}_{Y01}(1-\hat{\mathrm{PPV}}_{X1})\hat{\pi}^*_{11}}{\hat{\mathrm{PPV}}_{Y01}(1-\hat{\mathrm{PPV}}_{X1})\hat{\pi}^*_{11}+(1-\hat{\mathrm{NPV}}_{Y01})(1-\hat{\mathrm{PPV}}_{X0})\hat{\pi}^*_{10}}$$

$$\hat{\mathrm{SE}}_{Y00}=\frac{\hat{\mathrm{PPV}}_{Y00}\hat{\mathrm{NPV}}_{X1}\hat{\pi}^*_{01}}{\hat{\mathrm{PPV}}_{Y00}\hat{\mathrm{NPV}}_{X1}\hat{\pi}^*_{01}+(1-\hat{\mathrm{NPV}}_{Y00})\hat{\mathrm{NPV}}_{X0}\hat{\pi}^*_{00}}$$

$$\hat{\mathrm{SP}}_{Y11}=\frac{\hat{\mathrm{NPV}}_{Y11}\hat{\mathrm{PPV}}_{X0}\hat{\pi}^*_{10}}{\hat{\mathrm{NPV}}_{Y11}\hat{\mathrm{PPV}}_{X0}\hat{\pi}^*_{10}+(1-\hat{\mathrm{PPV}}_{Y11})\hat{\mathrm{PPV}}_{X1}\hat{\pi}^*_{11}}$$

$$\hat{\mathrm{SP}}_{Y10}=\frac{\hat{\mathrm{NPV}}_{Y10}(1-\hat{\mathrm{NPV}}_{X0})\hat{\pi}^*_{00}}{\hat{\mathrm{NPV}}_{Y10}(1-\hat{\mathrm{NPV}}_{X0})\hat{\pi}^*_{00}+(1-\hat{\mathrm{PPV}}_{Y10})(1-\hat{\mathrm{NPV}}_{X1})\hat{\pi}^*_{01}}$$

$$\hat{\mathrm{SP}}_{Y01}=\frac{\hat{\mathrm{NPV}}_{Y01}(1-\hat{\mathrm{PPV}}_{X0})\hat{\pi}^*_{10}}{\hat{\mathrm{NPV}}_{Y01}(1-\hat{\mathrm{PPV}}_{X0})\hat{\pi}^*_{10}+(1-\hat{\mathrm{PPV}}_{Y01})(1-\hat{\mathrm{PPV}}_{X1})\hat{\pi}^*_{11}}$$

$$\hat{\mathrm{SP}}_{Y00}=\frac{\hat{\mathrm{NPV}}_{Y00}\hat{\mathrm{NPV}}_{X0}\hat{\pi}^*_{00}}{\hat{\mathrm{NPV}}_{Y00}\hat{\mathrm{NPV}}_{X0}\hat{\pi}^*_{00}+(1-\hat{\mathrm{PPV}}_{Y00})\hat{\mathrm{NPV}}_{X1}\hat{\pi}^*_{01}}$$

$$\hat{\mathrm{SE}}_{X1}=\frac{\hat{\mathrm{PPV}}_{Y11}\hat{\mathrm{PPV}}_{X1}\hat{\pi}^*_{11}+(1-\hat{\mathrm{NPV}}_{Y11})\hat{\mathrm{PPV}}_{X0}\hat{\pi}^*_{10}}{\hat{\mathrm{PPV}}_{Y11}\hat{\mathrm{PPV}}_{X1}\hat{\pi}^*_{11}+(1-\hat{\mathrm{NPV}}_{Y11})\hat{\mathrm{PPV}}_{X0}\hat{\pi}^*_{10}+\hat{\mathrm{PPV}}_{Y10}(1-\hat{\mathrm{NPV}}_{X1})\hat{\pi}^*_{01}+(1-\hat{\mathrm{NPV}}_{Y10})(1-\hat{\mathrm{NPV}}_{X0})\hat{\pi}^*_{00}}$$

$$\hat{\mathrm{SE}}_{X0}=\frac{\hat{\mathrm{NPV}}_{Y11}\hat{\mathrm{PPV}}_{X0}\hat{\pi}^*_{10}+(1-\hat{\mathrm{PPV}}_{Y11})\hat{\mathrm{PPV}}_{X1}\hat{\pi}^*_{11}}{\hat{\mathrm{NPV}}_{Y11}\hat{\mathrm{PPV}}_{X0}\hat{\pi}^*_{10}+(1-\hat{\mathrm{PPV}}_{Y11})\hat{\mathrm{PPV}}_{X1}\hat{\pi}^*_{11}+\hat{\mathrm{NPV}}_{Y10}(1-\hat{\mathrm{NPV}}_{X0})\hat{\pi}^*_{00}+(1-\hat{\mathrm{PPV}}_{Y_{10}})(1-\hat{\mathrm{NPV}}_{X1})\hat{\pi}^*_{01}}$$

$$\hat{\mathrm{SP}}_{X1}=\frac{\hat{\mathrm{PPV}}_{Y00}\hat{\mathrm{NPV}}_{X1}\hat{\pi}^*_{01}+(1-\hat{\mathrm{NPV}}_{Y00})\hat{\mathrm{NPV}}_{X0}\hat{\pi}^*_{00}}{\hat{\mathrm{PPV}}_{Y00}\hat{\mathrm{NPV}}_{X1}\hat{\pi}^*_{01}+(1-\hat{\mathrm{NPV}}_{Y00})\hat{\mathrm{NPV}}_{X0}\hat{\pi}^*_{00}+\hat{\mathrm{PPV}}_{Y01}(1-\hat{\mathrm{PPV}}_{X1})\hat{\pi}^*_{11}+(1-\hat{\mathrm{NPV}}_{Y01})(1-\hat{\mathrm{PPV}}_{X0})\hat{\pi}^*_{10}}$$

$$\hat{\mathrm{SP}}_{X0}=\frac{\hat{\mathrm{NPV}}_{Y00}\hat{\mathrm{NPV}}_{X0}\hat{\pi}^*_{00}+(1-\hat{\mathrm{PPV}}_{Y00})\hat{\mathrm{NPV}}_{X1}\hat{\pi}^*_{01}}{\hat{\mathrm{NPV}}_{Y00}\hat{\mathrm{NPV}}_{X0}\hat{\pi}^*_{00}+(1-\hat{\mathrm{PPV}}_{Y00})\hat{\mathrm{NPV}}_{X1}\hat{\pi}^*_{01}+\hat{\mathrm{NPV}}_{Y01}(1-\hat{\mathrm{PPV}}_{X0})\hat{\pi}^*_{10}+(1-\hat{\mathrm{PPV}}_{Y01})(1-\hat{\mathrm{PPV}}_{X1})\hat{\pi}^*_{11}}$$

Appendix 3: closed-form ML estimators for πs with (X, Y, X*, Y*) available on all subjects

In general, $L_{\text{full}} = L_{\text{main}} \times L_{\text{val}}$. When ($X$, Y, X^*, Y^*) is measured on the whole sample, every subject can be regarded as a validation observation, so that there are no main study observations (i.e. $M = 0$ in Section 2.5) in this special case. Thus, $L_{\text{full}} = L_{\text{val}}$.

Under the most general misclassification model (Model 1 in Section 2.7), we may write the likelihood as follows:

$$\begin{aligned}
L_{\text{full}} &= L_{\text{val}} \\
&= \prod_{i=1}^{n_v} (\mathrm{SE}_{Y11}\mathrm{SE}_{X1}\pi_{11})^{x_i y_i x_i^* y_i^*} ((1-\mathrm{SP}_{Y11})\mathrm{SE}_{X0}\pi_{10})^{x_i(1-y_i)x_i^* y_i^*} (\mathrm{SE}_{Y01}(1-\mathrm{SP}_{X1})\pi_{01})^{(1-x_i)y_i x_i^* y_i^*} \\
&\times ((1-\mathrm{SP}_{Y01})(1-\mathrm{SP}_{X0})\pi_{00})^{(1-x_i)(1-y_i)x_i^* y_i^*} ((1-\mathrm{SE}_{Y11})\mathrm{SE}_{X1}\pi_{11})^{x_i y_i x_i^*(1-y_i^*)} (\mathrm{SP}_{Y11}\mathrm{SE}_{X0}\pi_{10})^{x_i(1-y_i)x_i^*(1-y_i^*)} \\
&\times ((1-\mathrm{SE}_{Y01})(1-\mathrm{SP}_{X1})\pi_{01})^{(1-x_i)y_i x_i^*(1-y_i^*)} (\mathrm{SP}_{Y01}(1-\mathrm{SP}_{X0})\pi_{00})^{(1-x_i)(1-y_i)x_i^*(1-y_i^*)} \\
&\times (\mathrm{SE}_{Y10}(1-\mathrm{SE}_{X1})\pi_{11})^{x_i y_i(1-x_i^*)y_i^*} ((1-\mathrm{SP}_{Y10})(1-\mathrm{SE}_{X0})\pi_{10})^{x_i(1-y_i)(1-x_i^*)y_i^*} (\mathrm{SE}_{Y00}\mathrm{SP}_{X1}\pi_{01})^{(1-x_i)y_i(1-x_i^*)y_i^*} \\
&\times ((1-\mathrm{SP}_{Y00})\mathrm{SP}_{X0}\pi_{00})^{(1-x_i)(1-y_i)(1-x_i^*)y_i^*} ((1-\mathrm{SE}_{Y10})(1-\mathrm{SE}_{X1})\pi_{11})^{x_i y_i(1-x_i^*)(1-y_i^*)} \\
&\times (\mathrm{SP}_{Y10}(1-\mathrm{SE}_{X0})\pi_{10})^{x_i(1-y_i)(1-x_i^*)(1-y_i^*)} ((1-\mathrm{SE}_{Y00})\mathrm{SP}_{X1}\pi_{01})^{(1-x_i)y_i(1-x_i^*)(1-y_i^*)} \\
&\times (\mathrm{SP}_{Y00}\mathrm{SP}_{X0}\pi_{00})^{(1-x_i)(1-y_i)(1-x_i^*)(1-y_i^*)}
\end{aligned}$$

The above term can be rewritten as:

$$
\begin{aligned}
L_{\text{full}} &= L_{\text{val}} \\
&= (\text{SE}_{Y11}\text{SE}_{X1}\pi_{11})^{\sum_{i=1}^{n_v} x_i y_i x_i^* y_i^*} ((1-\text{SP}_{Y11})\text{SE}_{X0}\pi_{10})^{\sum_{i=1}^{n_v} x_i (1-y_i) x_i^* y_i^*} (\text{SE}_{Y01}(1-\text{SP}_{X1})\pi_{01})^{\sum_{i=1}^{n_v} (1-x_i) y_i x_i^* y_i^*} \\
&\quad \times ((1-\text{SP}_{Y01})(1-\text{SP}_{X0})\pi_{00})^{\sum_{i=1}^{n_v} (1-x_i)(1-y_i) x_i^* y_i^*} ((1-\text{SE}_{Y11})\text{SE}_{X1}\pi_{11})^{\sum_{i=1}^{n_v} x_i y_i x_i^* (1-y_i^*)} \\
&\quad \times (\text{SP}_{Y11}\text{SE}_{X0}\pi_{10})^{\sum_{i=1}^{n_v} x_i (1-y_i) x_i^* (1-y_i^*)} ((1-\text{SE}_{Y01})(1-\text{SP}_{X1})\pi_{01})^{\sum_{i=1}^{n_v} (1-x_i) y_i x_i^* (1-y_i^*)} \\
&\quad \times (\text{SP}_{Y01}(1-\text{SP}_{X0})\pi_{00})^{\sum_{i=1}^{n_v} (1-x_i)(1-y_i) x_i^* (1-y_i^*)} (\text{SE}_{Y10}(1-\text{SE}_{X1})\pi_{11})^{\sum_{i=1}^{n_v} x_i y_i (1-x_i^*) y_i^*} \\
&\quad \times ((1-\text{SP}_{Y10})(1-\text{SE}_{X0})\pi_{10})^{\sum_{i=1}^{n_v} x_i (1-y_i)(1-x_i^*) y_i^*} (\text{SE}_{Y00}\text{SP}_{X1}\pi_{01})^{\sum_{i=1}^{n_v} (1-x_i) y_i (1-x_i^*) y_i^*} \\
&\quad \times ((1-\text{SP}_{Y00})\text{SP}_{X0}\pi_{00})^{\sum_{i=1}^{n_v} (1-x_i)(1-y_i)(1-x_i^*) y_i^*} ((1-\text{SE}_{Y10})(1-\text{SE}_{X1})\pi_{11})^{\sum_{i=1}^{n_v} x_i y_i (1-x_i^*)(1-y_i^*)} \\
&\quad \times (\text{SP}_{Y10}(1-\text{SE}_{X0})\pi_{10})^{\sum_{i=1}^{n_v} x_i (1-y_i)(1-x_i^*)(1-y_i^*)} ((1-\text{SE}_{Y00})\text{SP}_{X1}\pi_{01})^{\sum_{i=1}^{n_v} (1-x_i) y_i (1-x_i^*)(1-y_i^*)} \\
&\quad \times (\text{SP}_{Y00}\text{SP}_{X0}\pi_{00})^{\sum_{i=1}^{n_v} (1-x_i)(1-y_i)(1-x_i^*)(1-y_i^*)}
\end{aligned}
$$

which is $\pi_{11}^{\sum_{i=1}^{n_v} x_i y_i} \pi_{10}^{\sum_{i=1}^{n_v} x_i (1-y_i)} \pi_{01}^{\sum_{i=1}^{n_v} (1-x_i) y_i} \pi_{00}^{\sum_{i=1}^{n_v} (1-x_i)(1-y_i)}$ multiplied by a piece only involving misclassification probabilities (denoted by P). As a result,

$$
\log(L_{\text{full}}) = \log(P) + \sum_{i=1}^{n_v} x_i y_i \times \log(\pi_{11}) + \sum_{i=1}^{n_v} x_i (1-y_i) \times \log(\pi_{10}) + \sum_{i=1}^{n_v} (1-x_i) y_i \times \log(\pi_{01}) + \sum_{i=1}^{n_v} (1-x_i)(1-y_i) \times \log(\pi_{00}) \quad [6]
$$

Since the term P does not involve primary parameters, we can maximize the above log likelihood in terms of the πs easily with closed-form solutions as $\hat{\pi}_{ij} = \frac{I_{x=i, y=j}}{n_v}$, where I is defined similarly as in Section 2.5. The standard errors can be derived by taking the second derivatives of eq. [6] with respect to the πs. It should be noted that if only interested in primary parameters, the log likelihood expression in eq. [6] has exactly the same form when ignoring (X^*, Y^*). This confirms that inference on the πs stays the same no matter whether surrogate information is taken into account or not, when all participants in the study receive gold-standard evaluations. Under other less general misclassification models, the conclusion holds by following a similar argument.

Appendix 4: a summary of the fourth HERS visit data for models in Section 3

Table 7

BV and TRICH data of 916 participants at the fourth HERS visit

	Main study		
CLIN BV	**Wet mount TRICH**		**Total**
	−	+	
−	497	23	520
+	138	29	167
Total	635	52	687
Internal validation sample			
CLIN BV = 1, WET TRICH = 1, LAB BV = 1, CULTURE TRICH = 1		7	
CLIN BV = 1, WET TRICH = 1, LAB BV = 1, CULTURE TRICH = 0		0	
CLIN BV = 1, WET TRICH = 1, LAB BV = 0, CULTURE TRICH = 1		3	
CLIN BV = 1, WET TRICH = 1, LAB BV = 0, CULTURE TRICH = 0		0	
CLIN BV = 1, WET TRICH = 0, LAB BV = 1, CULTURE TRICH = 1		11	
CLIN BV = 1, WET TRICH = 0, LAB BV = 1, CULTURE TRICH = 0		28	
CLIN BV = 1, WET TRICH = 0, LAB BV = 0, CULTURE TRICH = 1		0	
CLIN BV = 1, WET TRICH = 0, LAB BV = 0, CULTURE TRICH = 0		8	
CLIN BV = 0, WET TRICH = 1, LAB BV = 1, CULTURE TRICH = 1		2	
CLIN BV = 0, WET TRICH = 1, LAB BV = 1, CULTURE TRICH = 0		0	
CLIN BV = 0, WET TRICH = 1, LAB BV = 0, CULTURE TRICH = 1		4	
CLIN BV = 0, WET TRICH = 1, LAB BV = 0, CULTURE TRICH = 0		1	
CLIN BV = 0, WET TRICH = 0, LAB BV = 1, CULTURE TRICH = 1		11	
CLIN BV = 0, WET TRICH = 0, LAB BV = 1, CULTURE TRICH = 0		34	
CLIN BV = 0, WET TRICH = 0, LAB BV = 0, CULTURE TRICH = 1		11	
CLIN BV = 0, WET TRICH = 0, LAB BV = 0, CULTURE TRICH = 0		109	
Total		229	

Table 1

Description and likelihood contributions for 16 possible types of observations under the internal validation sampling

Obs. type	Description	Likelihood contribution in terms of SE and SP	Likelihood contribution in terms of predictive values
1	$X^*=1, Y^*=1, X=1, Y=1$	$SE_{Y11}SE_{X1}\pi_{11}$	$PPV_{Y11}PPV_{X1}\ \pi_{11}^*$
2	$X^*=1, Y^*=1, X=1, Y=0$	$(1-SP_{Y11})SE_{X0}\pi_{10}$	$(1-PPV_{Y11})PPV_{X1}\ \pi_{11}^*$
3	$X^*=1, Y^*=1, X=0, Y=1$	$SE_{Y01}(1-SP_{X1})\pi_{01}$	$PPV_{Y01}(1-PPV_{X1})\ \pi_{11}^*$
4	$X^*=1, Y^*=1, X=0, Y=0$	$(1-SP_{Y01})(1-SP_{X0})\pi_{00}$	$(1-PPV_{Y01})(1-PPV_{X1})\ \pi_{11}^*$
5	$X^*=1, Y^*=0, X=1, Y=1$	$(1-SE_{Y11})SE_{X1}\pi_{11}$	$(1-NPV_{Y11})PPV_{X0}\ \pi_{10}^*$
6	$X^*=1, Y^*=0, X=1, Y=0$	$SP_{Y11}SE_{X0}\pi_{10}$	$NPV_{Y11}PPV_{X0}\pi_{10}^*$
7	$X^*=1, Y^*=0, X=0, Y=1$	$(1-SE_{Y01})(1-SP_{X1})\pi_{01}$	$(1-NPV_{Y01})(1-PPV_{X0})\ \pi_{10}^*$
8	$X^*=1, Y^*=0, X=0, Y=0$	$SP_{Y01}\ (1-SP_{X0})\pi_{00}$	$NPV_{Y01}(1-PPV_{X0})\ \pi_{10}^*$
9	$X^*=0, Y^*=1, X=1, Y=1$	$SE_{Y10}\ (1-SE_{X1})\pi_{11}$	$PPV_{Y10}(1-NPV_{X1})\ \pi_{01}^*$
10	$X^*=0, Y^*=1, X=1, Y=0$	$(1-SP_{Y10})(1-SE_{X0})\pi_{10}$	$(1-PPV_{Y10})(1-NPV_{X1})\ \pi_{01}^*$
11	$X^*=0, Y^*=1, X=0, Y=1$	$SE_{Y00}SP_{X1}\pi_{01}$	$PPV_{Y00}NPV_{X1}\pi_{01}^*$
12	$X^*=0, Y^*=1, X=0, Y=0$	$(1-SP_{Y00})SP_{X0}\pi_{00}$	$(1-PPV_{Y00})\ NPV_{X1}\pi_{01}^*$
13	$X^*=0, Y^*=0, X=1, Y=1$	$(1-SE_{Y10})(1-SE_{X1})\pi_{11}$	$(1-NPV_{Y10})\ (1-NPV_{X0})\ \pi_{00}^*$
14	$X^*=0, Y^*=0, X=1, Y=0$	$SP_{Y10}\ (1-SE_{X0})\pi_{10}$	$NPV_{Y10}(1-NPV_{X0})\pi_{00}^*$
15	$X^*=0, Y^*=0, X=0, Y=1$	$(1-SE_{Y00})SP_{X1}\pi_{01}$	$(1-NPV_{Y00})NPV_{X0}\pi_{00}^*$
16	$X^*=0, Y^*=0, X=0, Y=0$	$SP_{Y00}SP_{X0}\pi_{00}$	$NPV_{Y00}NPV_{X0}\pi_{00}^*$

Note: See Section 2.1 for the definitions of the terms.

Table 2

Results of analysis of 916 women at Visit 4 in the HERS, effects of correction models on OR estimates under various misclassification assumptions

Model	$\log(\hat{\mathrm{OR}})$ (StdErr)	$\hat{\mathrm{OR}}$ (95% CI)	AIC
Naïve[a]	1.54(0.26)	4.65 (2.81, 7.69)	
Gold standard[b]	1.14(0.18)	3.13 (2.21, 4.43)	
Main/internal validation: Model 1[c]	1.18(0.33)	3.24 (1.14, 5.35)	1,935.0
Main/internal validation: Model 2[d]	1.25(0.32)	3.48 (1.25, 5.71)	1,946.0
Main/internal validation: Model 3[e]	1.58(0.31)	4.84 (1.90, 7.78)	1,942.9

Notes:

[a] CLIN BV vs wet mount TRICH for all 916 subjects.

[b] LAB BV vs culture TRICH for all 916 subjects.

[c] 229 internal validation and 687 main study observations per simulation. Model 1 assumes dependent and differential misclassification.

[d] Model 2 assumes independent and differential misclassification.

[e] Model 3 assumes completely nondifferential misclassification.

Table 3

Results of simulations addressing main/internal validation study-based analysis mimicking HERS data

Model	$\log(\hat{\mathrm{OR}})_{(\mathrm{SD})}$	95% CI coverage
Naïve[a]	1.42 (0.23)	67.4%
Gold standard[b]	1.15 (0.18)	93.6%
Model 1[c]	1.16 (0.34)	95.7%
Model 2[d]	1.28 (0.34)	93.3%
Model 3[e]	1.58 (0.31)	72.4%

Notes: 500 simulations; 229 internal validation and 687 main study observations per simulation. True log(OR) = 1.14.

[a] $\hat{\mathrm{OR}}$ calculated using (Y^*, X^*) data.

[b] $\hat{\mathrm{OR}}$ calculated using (Y, X) data. $\mathrm{SE}_{x1} = 0.55$, $\mathrm{SP}_{x1} = 0.82$, $\mathrm{SE}_{x0} = 0.51$, $\mathrm{SP}_{x0} = 0.95$, $\mathrm{SE}_{y11} = 0.47$, $\mathrm{SP}_{y11} = 0.98$, $\mathrm{SE}_{y01} = 0.82$, $\mathrm{SP}_{y01} = 0.99$, $\mathrm{SE}_{y10} = 0.21$, $\mathrm{SP}_{y10} = 0.98$, $\mathrm{SE}_{y00} = 0.31$, and $\mathrm{SP}_{y00} = 0.99$.

[c] Model assuming dependent and differential misclassification.

[d] Model assuming independent and differential misclassification.

[e] Model assuming completely nondifferential misclassification.

Table 4

Performance of model selection with main/internal validation study-based analysis under a negative association

Model	$\log(\hat{\mathrm{OR}})_{(\mathrm{SD})}$	Mean SE	95% CI coverage
Setting 1: SE_X =0.60, SP_X = 0.90, SE_Y = 0.70, SP_Y = 0.80			
Naïve	−0.32 (0.15)	0.15	0
Gold standard	−1.10 (0.14)	0.15	95.4%
Model 1	−1.10 (0.28)	0.29	95.2%
Model 2	−1.10 (0.28)	0.28	94.8%
Model 3 (underlying model)	−1.10 (0.27)	0.27	95.4%
Model selection[a]	−1.10 (0.27)	0.27	95.4%
Setting 2: SE_{X1} = 0.60, SP_{X1} = 0.60, SE_{X0} = 0.90, SP_{X0} = 0.90, SE_{Y1} = 0.40, SP_{Y1} = 0.98, SE_{Y0} = 0.70, SP_{Y0} = 0.80			
Naïve	−0.61 (0.15)	0.15	9.4%
Gold standard	−1.10 (0.16)	0.15	93.2%
Model 1	−1.10 (0.30)	0.28	94.4%
Model 2 (underlying model)	−1.10 (0.29)	0.28	94.6%
Model 3	−1.28 (0.26)	0.26	90.0%
Model selection[b]	−1.10 (0.29)	0.28	94.2%
Setting 3: SE_{X1} = 0.60, SP_{X1} = 0.91, SE_{X0} = 0.48, SP_{X0} = 0.94, SE_{Y11} = 0.50, SP_{Y11} = 0.98, SE_{Y10} = 0.21, SP_{Y10} = 0.99, SE_{Y01} = 0.63, SP_{Y01} = 0.97, SE_{Y00} = 0.31, SP_{Y00} = 0.99			
Naïve	0.82 (0.27)	0.20	0
Gold standard	−1.11 (0.15)	0.15	94.6%
Model 1 (underlying model)	−1.12 (0.28)	0.27	94.1%
Model 2	−1.00 (0.27)	0.27	85.2%
Model 3	−0.62 (0.27)	0.27	58.3%
Model selection[c]	−1.11 (0.28)	0.27	93.2%

Notes: 500 simulation studies; 229 internal validation observations and 687 main study observations per simulation. Data were generated from a multinomial distribution with cell probabilities of (π_{11} = 0.10, π_{10} = 0.30, π_{01} = 0.30, π_{00} = 0.30). True log(OR) = −1.10. Naïve model uses (Y^*, X^*) data. Gold-standard model uses (Y, X) data. Model 1 assumes dependent and differential misclassification. Model 2 assumes independent and differential misclassification. Model 3 assumes completely nondifferential misclassification. Model selection based on the strategy described in Section 2.7.

[a] Model 3 selected 88.8% of the time.

[b] Model 2 selected 92.4% of the time.

[c] Model 1 selected 85.0% of the time.

Table 5

Performance of model selection with main/internal validation study-based analysis under a moderate positive association

Model	$\log(\hat{\mathrm{OR}})_{\mathbf{(SD)}}$	Mean SE	95% CI coverage
Setting 4: SE_X = 0.60, SP_X = 0.90, SE_Y = 0.70, SP_Y = 0.80			
Naïve	0.22(0.13)	0.14	1.2%
Gold standard	0.81(0.14)	0.14	94.6%
Model 1	0.82(0.27)	0.26	94.8%
Model 2	0.82(0.26)	0.26	95.0%
Model 3 (underlying model)	0.82(0.25)	0.25	95.8%
Model selection[a]	0.82(0.25)	0.25	95.8%
Setting 5: SE_{X1} = 0.60, SP_{X1} = 0.60, SE_{X0} = 0.90, SP_{X0} = 0.90, SE_{Y1} = 0.40, SP_{Y1} = 0.98, SE_{Y0} = 0.70, SP_{Y0} = 0.80			
Naïve	−0.28(0.14)	0.14	0
Gold standard	0.81(0.14)	0.14	94.6%
Model 1	0.81(0.26)	0.26	95.6%
Model 2 (underlying model)	0.81(0.25)	0.25	94.8%
Model 3	0.60(0.25)	0.25	84.6%
Model selection[b]	0.81(0.25)	0.25	95.0%
Setting 6: SE_{X1} = 0.60, SP_{X1} = 0.91, SE_{X0} = 0.48, SP_{X0} = 0.94, SE_{Y11} = 0.50, SP_{Y11} = 0.98, SE_{Y10} = 0.21, SP_{Y10} = 0.99, SE_{Y01} = 0.63, SP_{Y01} = 0.97, SE_{Y00} = 0.31, SP_{Y00} = 0.99			
Naïve	1.64(0.17)	0.17	7.2%
Gold standard	0.82(0.14)	0.14	94.6%
Model 1 (underlying model)	0.81(0.25)	0.26	95.0%
Model 2	0.93(0.24)	0.25	86.6%
Model 3	1.60(0.24)	0.24	17.0%
Model selection[c]	0.81(0.25)	0.26	94.4%

Notes: 500 simulation studies; 229 internal validation observations and 687 main study observations per simulation. Data were generated from a multinomial distribution with cell probabilities of (π_{11} = 0.30, π_{10} = 0.20, π_{01} = 0.20, π_{00} = 0.30). True log(OR) = 0.81. Naïve model uses (Y^*, X^*) data. Gold-standard model uses (Y, X) data. Model 1 assumes dependent and differential misclassification. Model 2 assumes independent and differential misclassification. Model 3 assumes completely nondifferential misclassification. Model selection based on the strategy described in Section 2.7.

[a] Model 3 selected 88.0% of the time.

[b] Model 2 selected 94.0% of the time.

[c] Model 1 selected 94.8% of the time.

Table 6

Performance of model selection with main/internal validation study-based analysis under a strong positive association

Model	$\log(\hat{\mathrm{OR}})_{(\mathrm{SD})}$	Mean SE	95% CI coverage
Setting 7: SE_X = 0.60, SP_X = 0.90, SE_Y = 0.70, SP_Y = 0.80			
Naïve	0.46(0.14)	0.14	0
Gold standard	1.80(0.14)	0.15	96.4%
Model 1	1.82(0.28)	0.30	96.8%
Model 2	1.82(0.28)	0.29	96.8%
Model 3 (underlying model)	1.82(0.27)	0.28	96.4%
Model selection[a]	1.82(0.28)	0.28	96.4%
Setting 8: SE_{X1} = 0.60, SP_{X1} = 0.60, SE_{X0} = 0.90, SP_{X0} = 0.90, SE_{Y1} = 0.40, SP_{Y1} = 0.98, SE_{Y0} = 0.70, SP_{Y0} = 0.80			
Naïve	−0.20(0.15)	0.15	0
Gold standard	1.80(0.16)	0.15	93.8%
Model 1	1.81(0.31)	0.29	93.6%
Model 2 (underlying model)	1.81(0.31)	0.29	93.4%
Model 3	1.59(0.30)	0.29	85.8%
Model selection[b]	1.81(0.31)	0.29	93.2%
Setting 9: SE_{X1} = 0.60, SP_{X1} = 0.91, SE_{X0} = 0.48, SP_{X0} = 0.94, SE_{Y11} = 0.50, SP_{Y11} = 0.98, SE_{Y10} = 0.21, SP_{Y10} = 0.99, SE_{Y01} = 0.63, SP_{Y01} = 0.97, SE_{Y00} = 0.31, SP_{Y00} = 0.99			
Naïve	1.98(0.18)	0.18	68.2%
Gold standard	1.79(0.14)	0.15	96.8%
Model 1 (underlying model)	1.80(0.28)	0.29	97.0%
Model 2	1.95(0.28)	0.28	92.8%
Model 3	2.57(0.27)	0.27	19.8%
Model selection[c]	1.80(0.28)	0.29	97.0%

Notes: 500 simulation studies; 229 internal validation observations and 687 main study observations per simulation. Data were generated from a multinomial distribution with cell probabilities of (π_{11} = 0.30, π_{10} = 0.10, π_{01} = 0.20, π_{00} = 0.40). True log (OR) = 1.79. Naïve model uses (Y^*, X^*) data. Gold-standard model uses (Y, X) data. Model 1 assumes dependent and differential misclassification. Model 2 assumes independent and differential misclassification. Model 3 assumes completely nondifferential misclassification. Model selection based on the strategy described in Section 2.7.

[a] Model 3 selected 87.2% of the time.

[b] Model 2 selected 90.6% of the time.

[c] Model 1 selected 95.8% of the time.