Published in final edited form as:
Epidemiology. 2011 July ; 22(4): 589–597. doi:10.1097/EDE.0b013e3182117c85.

Validation Data-Based Adjustments for Outcome Misclassification in Logistic Regression:An Illustration

Robert H. Lyles[1], **Li Tang**[1], **Hillary M. Superak**[1], **Caroline C. King**[2], **David D. Celentano**[3], **Yungtai Lo**[4], and **Jack D. Sobel**[5]

[1]Department of Biostatistics and Bioinformatics, The Rollins School of Public Health of Emory University, Atlanta, GA

[2]Centers for Disease Control and Prevention, Atlanta, GA

[3]Johns Hopkins Bloomberg School of Public Health, Baltimore, MD

[4]Montefiore Medical Center and Albert Einstein College of Medicine, Bronx, NY

[5]Wayne State University School of Medicine, Detroit, MI

Abstract

Misclassification of binary outcome variables is a known source of potentially serious bias when estimating adjusted odds ratios. Although researchers have described frequentist and Bayesian methods for dealing with the problem, these methods have seldom fully bridged the gap between statistical research and epidemiologic practice. In particular, there have been few real-world applications of readily grasped and computationally accessible methods that make direct use of internal validation data to adjust for differential outcome misclassification in logistic regression. In this paper, we illustrate likelihood-based methods for this purpose that can be implemented using standard statistical software. Using main study and internal validation data from the HIV Epidemiology Research Study, we demonstrate how misclassification rates can depend on the values of subject-specific covariates, and illustrate the importance of accounting for this dependence. Simulation studies confirm the effectiveness of the maximum likelihood approach. We emphasize clear exposition of the likelihood function itself, to permit the reader to easily assimilate appended computer code that facilitates sensitivity analyses as well as the efficient handling of main/external and main/internal validation-study data. These methods are readily applicable under random cross-sectional sampling, and we discuss the extent to which the main/internal analysis remains appropriate under outcome-dependent (case-control) sampling.

The consequences of misclassified binary outcome or exposure variables when estimating a crude odds ratio (OR) are well understood.[1–5] Existing literature also covers the use of validation data to estimate crude ORs while adjusting for misclassification in case-control and cross-sectional studies,[6–11] considering the relative merits of external versus internal validation study designs.[1; 11–12] In regression applications, many researchers advocate the use of validation data to adjust for measurement error in continuous predictors.[13–17]

Address for correspondence: Robert H. Lyles; Department of Biostatistics and Bioinformatics, The Rollins School of Public Health of Emory University, 1518 Clifton Rd. N.E., Atlanta, GA 30322 (phone: 404-727-1310; fax: 404-727-1370; rlyles@sph.emory.edu).

The findings and conclusions in this report are those of the authors and do not necessarily represent the official position of the Centers for Disease Control and Prevention.

Regarding outcome misclassification for discrete responses, Magder and Hughes[18] outline the problem under logistic regression and advocate maximum likelihood via an expectation-maximization algorithm.[19] Their work primarily addresses the case of known misclassification probabilities (i.e., sensitivities and specificities) characterizing the observed outcome variable. While continuing to focus on the known sensitivities/specificities case, Neuhaus[20] provides further insight into asymptotic bias and efficiency in the broader realm of the generalized linear model, as well as a more efficient computational maximum likelihood approach. Recent articles in the epidemiologic literature demonstrate Monte Carlo-based techniques that similarly facilitate sensitivity analyses with misclassified binary variables.[21–22]

Other related research includes extensions to settings with count or discrete survival outcomes.[23–25] To incorporate validation data, some authors gravitate toward Bayesian approaches using prior assumptions about misclassification probabilities.[26–28] From the parametric frequentist perspective, Carroll et al.[11] provide general expressions for likelihood functions that accommodate internal validation data. Alternative developments include robust modeling of sensitivity and specificity via kernel smoothers,[29] with comparisons of that approach versus parametrically modeling their dependence upon covariates.[30]

Our aim is to provide guidance for epidemiologists seeking accessible and efficient methods for obtaining validation data-based estimates of logistic regression parameters when the outcome is misclassified. We keep to a likelihood-based approach, as it avoids explicit specification of prior distributions and is readily facilitated for binary outcomes. In the general case, we model the dependence of sensitivity and specificity upon covariates via a second logistic regression model, promoting a flexible and intuitively appealing analytic approach.

The methodology that we illustrate is a direct expansion of the known misclassification rate setting considered by Magder and Hughes[18] and Neuhaus,[20] a covariate-adjusted extension of well-discussed methods for estimating crude ORs,[6–9] and ultimately an application of the general main/validation study maximum likelihood approach outlined in Carroll et al.[11] However, there have been few if any real-world applications of the latter approach making use of internal validation data, and such application presents computational challenges to the practicing epidemiologist. Thus, our goal is to bring this approach for addressing outcome misclassification in regression closer to the forefront of epidemiologic research. We pursue this aim by highlighting an instructive example involving misclassified outcome status in the HIV Epidemiology Research Study, by transparent exposition of appropriate likelihood functions, and by providing appendices with straightforward computer code that connects directly with that exposition.

METHODS

Assume we wish to fit the following logistic regression model to cross-sectional data (we discuss implications of case-control sampling later):

$$\tau = \text{logit}\left[\Pr\left(Y=1|X_1, X_2, \ldots, X_p\right)\right] = \beta_0 + \sum_{p=1}^{P} \beta_p X_p. \quad (1)$$

We use the symbol τ for easy reference to Eq. (1) in Appendix 1. Instead of the true (0,1) response Y, suppose the primary (main) study relies upon an error-prone (0,1) alternative Y*. It is known that misclassification in Y* potentially invalidates estimates of $(\beta_0,\ldots,\beta_P)$ based on the "naïve" model that replaces Y by Y* in Eq. (1). The magnitudes and directions

of biases in the naïve estimates depend upon the diagnostic properties of Y* as a substitute for Y.[18]

In the non-differential case,[1] the critical diagnostic properties boil down to two parameters, sensitivity (SE) and specificity (SP):

$$\mathrm{SE}=\Pr(\mathrm{Y}^*=1|\mathrm{Y}=1) \quad \text{and} \quad \mathrm{SP}=\Pr(\mathrm{Y}^*=0|\mathrm{Y}=0). \tag{2}$$

If misclassification is differential, however, then sensitivity and specificity can vary according to subject-specific variables, making effects of misclassification less predictable.[1] Thus, we define

$$\mathrm{SE}_{\mathbf{x}}=\Pr(\mathrm{Y}^*=1|\mathrm{Y}=1,\mathbf{X}=\mathbf{x}) \quad \text{and} \quad \mathrm{SP}_{\mathbf{x}}=\Pr(\mathrm{Y}^*=0|\mathrm{Y}=0,\mathbf{X}=\mathbf{x}), \tag{3}$$

where the vector $\mathbf{X}$ is usually some subset of $(X_1, X_2,\ldots, X_p)$.

Sensitivity Analyses

Suppose first that no validation data are available so that one has only main study data consisting of $(y_i^*, x_{i1}, \ldots, x_{iP})$ on the i^{th} experimental unit $(i=1,\ldots, n_m)$. In this case, each independent record contributes the following likelihood term:

$$L_{\mathrm{mi}}=\Pr(Y_i^*=y_i^*|\mathbf{X}=x_i)=\sum_{y_i=0}^{1}\Pr(Y_i^*=y_i^*|Y_i=y_i,\mathbf{X}=x_i)\Pr(Y_i=y_i|\mathbf{X}=x_i). \tag{4}$$

The first term after the summation in Eq. (4) is determined by SE_x and SP_x, while the second follows directly from Eq. (1). The overall likelihood is proportional to the product, i.e.,

$$L_{\mathrm{m}}=\prod_{\mathrm{i}=1}^{n_m}L_{\mathrm{mi}}. \tag{5}$$

While it may technically be possible to estimate $(\beta_1,\ldots,\beta_P)$ based only on main study data without supplying values of misclassification probabilities,[11; 18] these parameters will be weakly identifiable at best.[11] Neuhaus[20] notes that estimability of misclassification rates is compromised under mis-specification of the primary model [Eq. (1) here], further emphasizing the limited utility of a main study-only analysis. Thus, use of Eq. (5) is effectively limited to sensitivity analysis, wherein one supplies assumed values of SE_x and SP_x.

Both the EM approach of Magder and Hughes[18] and the alternative maximum likelihood conceptualization of Neuhaus[20] purport to maximize Eq. (5) after pre-specifying sensitivity and specificity values. For an implementation under non-differential misclassification that adapts readily to the differential case, Appendix 1 provides ready-to-use computer code utilizing the capacity for user-specified log-likelihood functions in the SAS NLMIXED procedure.[31] To specify the likelihood to the level of detail required for programming, note that Eq. (5) may be written as follows under non-differentiality:

$$\begin{aligned}L_{\mathrm{m}}=\prod_{\mathrm{i}=1}^{n_m}\{&[(1-\mathrm{SP})\times\Pr(\mathrm{Y}=0|\mathbf{X}=x_i)+\mathrm{SE}\times\Pr(\mathrm{Y}=1|\mathbf{X}=x_i)]^{y_i^*}\\&\times[\mathrm{SP}\times\Pr(\mathrm{Y}=0|\mathbf{X}=x_i)+(1-\mathrm{SE})\times\Pr(\mathrm{Y}=1|\mathbf{X}=x_i)]^{(1-y_i^*)},\end{aligned} \tag{6}$$

where $\Pr(Y = 1 | \mathbf{X}_i = \mathbf{x}_i) = \exp(\tau_i) / [1 + \exp(\tau_i)]$, with $\tau_i = \beta_0 + \sum_{p=1}^{P} \beta_p x_{ip}$ via Eq. (1). A special case of general likelihood expressions provided in Carroll et al.,[11] this structure is directly reflected in the first sample program in Appendix 1.

Main Study + External Validation Data: Non-differential Misclassification

Because sensitivity analysis is seldom a fully satisfying solution, we emphasize using validation data to estimate $(\beta_1, \ldots, \beta_P)$ in Eq. (1) without pre-specifying sensitivity and specificity values. When the validation sample is external[1] (i.e., separate from the main study), we confine attention to the non-differential case because external studies seldom measure the same covariates as the main study. External validation is also limited by a need to assume "transportability," i.e., that sensitivity and specificity parameters targeted in the validation sample are identical to those operating in the main study.[11–12] In the remainder of the paper, we use the shorthand "main/external" and "main/internal" to refer to settings in which main study data are combined with external or internal validation data, respectively.

Given that our primary focus is upon the analysis of main/internal study data as required in the motivating example, we relegate details of the main/external case to Appendix 2. The structure of the resulting main/external likelihood is reflected in the second SAS NLMIXED program found in Appendix 1.

Main Study + Internal Validation Data: Differential Misclassification

Our main interest lies in the case in which an internal validation sample (of size n_v) is randomly selected from the overall study sample. Again, main study experimental units contribute records of the form $(y_i^*, x_{i1}, \ldots, x_{iP})$. In contrast, resources are expended toward those selected for validation to augment their records with the true outcome status (y_i). Benefits of this supplemental data collection effort include removal of concern about transportability and flexibility to allow general patterns of differential misclassification.

As a first example, consider the case of two covariates, one continuous (X_1) and one binary (X_2), where sensitivity and specificity depend on X_2. That is, define $SE_t = \Pr(Y^*=1 \mid Y=1, X_2=t)$ and $SP_t = \Pr(Y^*=0 \mid Y=0, X_2=t)$ $(t=0,1)$. Main study contributions remain of the form in (4), yielding the following main study likelihood:

$$L_m = \prod_{i=1}^{n_m} \prod_{t=0}^{1} \left\{ [(1 - SP_t) \times \Pr(Y=0 | X_1=x_{1i}, X_2=t) + SE_t \times \Pr(Y=1 | X_1=x_{1i}, X_2=t)]^{y_i^*} \times [SP_t \times \Pr(Y=0 | X_1=x_{1i}, X_2=t) + (1 - SE_t) \times \Pr(Y=1 | X_1=x_{1i}, X_2=t)]^{(1-y_i^*)} \right\}^{I(x_{2i}=t)}, \quad (7)$$

where I(.) is a binary (0,1) indicator for whether the condition in parentheses is true. In contrast, internal validation data records contribute terms of the form

$$L_{vj} = \Pr\left(Y_j^* = y_j^*, Y_j = y_j | \mathbf{X} = x_i\right) = \Pr\left(Y_j^* = y_j^* | Y_j = y_j, \mathbf{X} = x_i\right) \Pr\left(Y_j = y_j | \mathbf{X} = x_i\right),$$

yielding an internal validation subsample likelihood as follows:

$$L_{\mathrm{v}}=\prod_{\mathrm{j}=1}^{n_v}\prod_{\mathrm{t}=0}^{1}\left\{\left[\mathrm{SE_t}\times\Pr\left(\mathrm{Y}=1|\mathrm{X}_1=\mathrm{x}_{1\mathrm{j}},\mathrm{X}_2=\mathrm{t}\right)\right]\mathrm{y}_{\mathrm{j}}^{*}y_j\times\left[(1-\mathrm{SP_t})\times\Pr\left(\mathrm{Y}=0|\mathrm{X}_1=\mathrm{x}_{1\mathrm{j}},\mathrm{X}_2=\mathrm{t}\right)\right]\mathrm{y}_{\mathrm{j}}^{*}\left(1-y_j\right)\right.$$
$$\left.\times\left[(1-\mathrm{SE_t})\times\Pr\left(\mathrm{Y}=1|\mathrm{X}_1=\mathrm{x}_{1\mathrm{j}},\mathrm{X}_2=\mathrm{t}\right)\right]\left(1-\mathrm{y}_{\mathrm{j}}^{*}\right)y_j\times\left[\mathrm{SP_t}\times\Pr\left(\mathrm{Y}=0|\mathrm{X}_1=\mathrm{x}_{1\mathrm{j}},\mathrm{X}_2=\mathrm{t}\right)\right]\left(1-\mathrm{y}_{\mathrm{j}}^{*}\right)\left(1-y_j\right)\right\}\mathrm{I}\left(\mathrm{x}_{2\mathrm{j}}=\mathrm{t}\right). \quad (8)$$

Again, the full likelihood is proportional to $L = L_{\mathrm{m}}\times L_{\mathrm{v}}$.

For the general case in which model (1) includes arbitrary predictors ($X_1,\ldots, X_P$), we assume sensitivity and specificity depend on ($X_2^*,\ldots, X_K^*$), which may denote a subset of ($X_1,\ldots, X_P$) and/or include other variables or interaction terms. We favor a second logistic model to define associations between these predictors and sensitivity / specificity:

$$\eta_{\mathrm{y}}=\mathrm{logit}\left[\Pr\left(\mathrm{Y}^{*}=1|\mathrm{Y}=\mathrm{y},\mathrm{X}_2^{*},\mathrm{X}_3^{*},\ldots,\mathrm{X}_{\mathrm{k}}^{*}\right)\right]=\theta_0+\theta_1\mathrm{y}+\sum_{\mathrm{k}=2}^{\mathrm{K}}\theta_{\mathrm{k}}\mathrm{X}_{\mathrm{k}}^{*}\ \ (\mathrm{y}=0,1). \quad (9)$$

Assuming an adequate internal validation sample, Eq. (9) allows us to flexibly account for differential misclassification. It does so in a potentially robust manner when ($X_2^*,\ldots X_K^*$) consists of categorical variables. For subject i contributing predictor values $\mathbf{x}_{\mathrm{i}}$, Eq. (9) implies that

$$\mathrm{SE}_{x_i}=\Pr\left(\mathrm{Y}^{*}=1|\mathrm{Y}=1,\mathbf{X}=x_i\right)=\frac{\exp\left(\eta_{\mathrm{i}1}\right)}{1+\exp\left(\eta_{\mathrm{i}1}\right)}$$

and

$$\mathrm{SP}_{x_i}=\Pr\left(\mathrm{Y}^{*}=0|\mathrm{Y}=0,\mathbf{X}=x_i\right)=1-\frac{\exp\left(\eta_{\mathrm{i}0}\right)}{1+\exp\left(\eta_{\mathrm{i}0}\right)}, \quad (10)$$

where $\eta_{\mathrm{iy}}=\theta_0+\theta_1\mathrm{y}+\sum_{\mathrm{k}=2}^{\mathrm{K}}\theta_{\mathrm{k}}\mathrm{x}_{\mathrm{ik}}^{*}$. Maximum likelihood estimates (MLEs) for the differential sensitivity and specificity parameters follow from the MLE of $\boldsymbol{\theta}=(\theta_1,\ldots,\theta_{\mathrm{K}})$.

The full likelihood for this general case is proportional to $L = L_{\mathrm{m}}\times L_{\mathrm{v}}$, where

$$L_{\mathrm{m}}=\prod_{\mathrm{i}=1}^{n_m}\left\{\left[\left(1-\mathrm{SP}_{x_i}\right)\times\Pr\left(\mathrm{Y}=0|\mathbf{X}=x_i\right)+\mathrm{SE}_{x_i}\times\Pr\left(\mathrm{Y}=1|\mathbf{X}=x_i\right)\right]^{y_i^{*}}\right.$$
$$\left.\times\left[\mathrm{SP}_{x_i}\times\Pr\left(\mathrm{Y}=0|\mathbf{X}=x_i\right)+\left(1-\mathrm{SE}_{x_i}\right)\times\Pr\left(\mathrm{Y}=1|\mathbf{X}=x_i\right)\right]^{\left(1-y_i^{*}\right)}\right. \quad (11)$$

(identical to Eq. (6) except for covariate effects on sensitivity and specificity), and

$$L_{\mathrm{v}}=\prod_{\mathrm{j}=1}^{n_v}\left\{\left[\mathrm{SE}_{\mathbf{x}_{\mathrm{j}}}\times\Pr\left(\mathrm{Y}=1|\mathbf{X}=\mathbf{x}_{\mathrm{j}}\right)\right]^{\mathrm{y}_{\mathrm{j}}^{*}y_j}\times\left[\left(1-\mathrm{SP}_{\mathbf{x}_{\mathrm{j}}}\right)\times\Pr\left(\mathrm{Y}=0|\mathbf{X}=\mathbf{x}_{\mathrm{j}}\right)\right]^{\mathrm{y}_{\mathrm{j}}^{*}\left(1-y_j\right)}\right.$$
$$\left.\times\left[\left(1-\mathrm{SE}_{\mathbf{x}_{\mathrm{j}}}\right)\times\Pr\left(\mathrm{Y}=1|\mathbf{X}=\mathbf{x}_{\mathrm{j}}\right)\right]^{\left(1-\mathrm{y}_{\mathrm{j}}^{*}\right)y_j}\times\left[\mathrm{SP}_{\mathbf{x}_{\mathrm{j}}}\times\Pr\left(\mathrm{Y}=0|\mathbf{X}=\mathbf{x}_{\mathrm{j}}\right)\right]^{\left(1-\mathrm{y}_{\mathrm{j}}^{*}\right)\left(1-y_j\right)}. \quad (12)$$

This likelihood structure is reflected in the third SAS NLMIXED program in Appendix 1 (http://links.lww.com). The likelihood itself is equivalent to a general expression found in the paper by Carroll et al.[11] We present it more explicitly here to enhance its clarity and connection with the provided program.

As with any parametric model, Eq. (9) makes likelihood ratio tests available based on Eq. (11)–(12) to aid in model selection and assess whether predictors are associated with sensitivity and specificity. This permits testing the hypothesis of completely non-differential misclassification, i.e., H_0: $(\theta_2 = \theta_3 = \ldots = \theta_K) = 0$.

Comments Regarding Case-Control Data

Prior treatments of outcome misclassification[18] offered limited or no applicability under outcome-dependent sampling, despite well-known classical results[32] establishing the utility of logistic regression for retrospective studies. It is thus of interest to explore whether and to what extent the recommended maximum likelihood approach accommodates the case-control design. By "case-control" here, we imply that sampling is done based on the error-prone response (Y*), with a higher sampling probability applied to "cases" (those with Y*=1) than to "controls" (those with Y*=0). We find that, with certain caveats, the internal validation study-based analysis proposed here can be used without modification despite the application of such "case" oversampling.

Specifically, the method described in the previous subsection yields valid estimates of $(\beta_1, \ldots, \beta_P)$under model (1) when sampling favors those with Y*=1, assuming non-differential misclassification of case/control status. As with the classic case,[32] the intercept loses its original interpretation. For similar reasons, the likelihood-based estimates of sensitivity and specificity will no longer reflect the true diagnostic properties of Y*. Rather, these tend to be inflated and deflated, respectively, in concert with the oversampling of cases according to Y*. In fact, the fallibility of the internal validation-based sensitivity/specificity estimators due to "case" oversampling is key to the validity of the $(\beta_1, \ldots, \beta_P)$ estimates, as these estimators reflect the "operating" sensitivity and specificity of Y* under the sampling strategy employed. In contrast, direct analysis based on external validation data (or even employing correct assumed values of sensitivity and specificity) misconstrues the "operating" sensitivity and specificity, generally yielding inconsistent estimates of $(\beta_1, \ldots, \beta_P)$. This may explain why methods[18; 20–22] that are not based on internal validation data encounter problems for case-control studies.

The validity of the main/internal validation study-based maximum likelihood approach for such case-control sampling with non-differential outcome misclassification recalls theoretical results in the statistical literature,[33] and can be demonstrated by noting that terms involving the selection probabilities applied to those with Y*=1 and Y*=0 factor out of the likelihood. In contrast, no such clean factorization occurs under differential misclassification. Nevertheless, if differential outcome misclassification is appropriately modeled via Eq. (11)–(12), empirical evidence via simulation under large samples suggests that the MLEs for some elements of $(\beta_1, \ldots, \beta_P)$ in model (1) may remain valid under "case" oversampling. Specifically, our experimentation suggests that β coefficients in model (1) remain reliably estimable if they correspond to predictor variables that are not needed in the second regression model (9) that defines sensitivity and specificity. A simulation study illustrating these points follows after the example section.

EXAMPLE

Our example concerns data on bacterial vaginosis status for women in the HIV Epidemiology Research Study. A total of 1,310 (871 HIV-infected and 439 at-risk uninfected) women were enrolled into this prospective study across four U.S. cities from 1993 to 1995.[34] Researchers diagnosed bacterial vaginosis semi-annually by two different techniques, referred to as the "CLIN" (clinically-based) and "LAB" (laboratory-based) methods. A CLIN diagnosis required the presence of three or more specific clinical conditions based on a modification of Amsel's criteria,[35] while LAB diagnoses were made

via a sophisticated Gram-staining technique.[36] Prior references[37–38] provide details on these methods in the study. As in Gallo et al.,[38] we treat the more costly LAB method as a gold standard assessment, while the CLIN approach represents an accessible error-prone substitute. These authors found evidence of low sensitivity for the CLIN method, and suggested that its accuracy may suffer due to wide heterogeneity in bacterial vaginosis cases or due to the need for technicians to be trained in order to properly apply the subjective Amsel criteria.[38]

A unique feature of this example is that both LAB and CLIN diagnoses were made regularly. Thus, in addition to fitting a "naïve" main study-only version of model (1) with CLIN status (Y*) substituted for LAB (Y), we were able to fit Eq. (1) to data using the assumed gold standard (Y) on all subjects. While the illustration of validation data-based adjusted analyses then requires ignoring LAB data on a random subset, an advantage is that we have an "ideal" complete-data model for comparison.

We use data from the 4th semi-annual study visit on 982 black, white, and Hispanic women who were 25 years or older at enrollment. Available variables potentially associated with bacterial vaginosis status include age, race, HIV status (0 if negative, 1 if positive), and HIV risk group (0 if via sexual contact; 1 if intravenous drug use). Study site and CD4 counts among HIV positives showed little association with bacterial vaginosis status in this sample.

Median age at enrollment was 37 years. Other potential bacterial vaginosis risk factors are distributed as follows: race/ethnicity (60% black, 24% white, 16% Hispanic); HIV status (69% positive, 31% negative); HIV risk group (47% sexual, 53% intravenous drug use). Among women with data on bacterial vaginosis, 41% were positive via the LAB method, versus 25% based on CLIN. Unadjusted estimates were 0.53 (sensitivity) and 0.94 (specificity), suggesting that CLIN yields a low risk of false positives but high risk of false negatives.

For an "ideal" comparative analysis, we first fit Eq. (1) to all women, with the gold standard diagnosis (LAB; 1 vs. 0) as the outcome. Preliminary analyses revealed similar bacterial vaginosis prevalence among white and Hispanic women, so we created a binary variable (0 if non-black, 1 if black). Initially dichotomizing age at the median, we assessed second- and higher-order interactions among age, race, HIV status, and risk group. A likelihood ratio test supported elimination of all 11 interaction terms.

A total of 924 women, with complete data on both bacterial vaginosis assessments and all risk factors, contributed to the fitted models summarized in Table 1. The upper half of the table summarizes the fit of the resulting version of model (1) for LAB status, in which we treat age (in years) continuously:

$$\text{logit}\left[\Pr(\text{LAB}=1)\right]=\beta_0+\beta_1\text{BLACK}+\beta_2\text{RISKGRP}+\beta_3\text{HIVPOS}+\beta_4\text{AGE}. \quad (13)$$

We then fit the same model upon substituting the error-prone CLIN diagnosis as the outcome (lower half of Table 1). The two analyses differ markedly in terms of magnitude of the estimated OR for HIV risk group (1.50 for LAB, 2.68 for CLIN), and directionality of the estimated OR for HIV status (1.19 for LAB, 0.71 for CLIN).

To illustrate misclassification adjustment, we selected a random internal validation subset of size n_V=300 women. Predictor selection via model (9) fit to these 300 women revealed no independent association between race and CLIN status. Pairwise and higher-order interactions among LAB status, risk group, HIV status, and age (dichotomized for purposes of estimating sensitivity and specificity) were non-significant as a group. The version of Eq. (9) utilized in the main/internal validation study likelihood is

$$\text{logit}\left[\Pr(\text{CLIN}=1)\right]=\theta_0+\theta_1\text{LAB}+\theta_2\text{RISKGRP}+\theta_3\text{HIVPOS}+\theta_4\text{AGEGTMED}, \quad (14)$$

where AGEGTMED indicates whether a subject's age at enrollment exceeded the median.

The upper half of Table 2 summarizes a complete analysis of the data via the joint likelihood in Eq. (11)–(12). For comparison, the lower half of Table 2 gives corresponding results assuming non-differential misclassification [restricting $\theta_2=\theta_3=\theta_4=0$ in Eq. (14)]. The likelihood ratio test comparing the joint models with and without the non-differentiality assumption was highly significant ($\chi^2=20.1$, $P<0.001$), strongly confirming a need to account for dependence of the sensitivity and specificity of the CLIN diagnosis upon subject-specific covariates. Note that the analysis in the upper half of Table 2 yields the same interpretations as the "ideal" analysis (upper half, Table 1), in terms of directionalities and magnitudes of the estimated ORs. In contrast, results in the lower half of Table 2 are similar to those of the "naïve" analysis (lower half, Table 1), showing an elevated estimate for risk group and negative directionality for HIV status. This highlights the value of internal validation data for modeling sensitivity and specificity.

Table 3 provides the MLE of (θ_0, θ_1, θ_2, θ_3, θ_4) in Eq. (14) based on the joint likelihood Eq. (11)–(12). Note that all three predictors (risk group, HIV status, and age) are independently associated with sensitivity and specificity. Table 3 provides corresponding MLEs of (SE, SP) via equations. (9)–(10), with multivariate delta method-based standard errors (details available from the authors). Holding other variables constant, sensitivity tends to be higher (and specificity lower) for those who are in the intravenous drug use risk group, younger, or HIV-negative. The variations in these estimates give further credence to the differential nature of outcome misclassification in this real-data example.

SIMULATION STUDIES

Simulation I: Mimicking Real-Data Example

Our primary simulation experiment evaluates the general main/internal validation study analysis outlined in (9)–(12), under conditions mimicking the example. Four predictors (X_1–X_4) were randomly generated with distributions like those observed at study Visit 4 for "Race" (black vs. non-black), "Risk Group," "HIV status," and "age," respectively. True outcomes (Y) were simulated according to Eq. (13), with β coefficients equal to the estimates reported in the top portion of Table 2. Error-prone outcomes (Y*) were generated via Eq. (14), with θ's equal to the estimates at the top of Table 3. For 1,000 such datasets, we conducted the "naïve" analysis in addition to two main/internal validation analyses based on Eq. (11)–(12). The first of these assumed the appropriate differential misclassification model, and the second incorrectly assumed non-differentiality.

Table 4 summarizes the results. The "naïve" analysis produces highly biased estimates, with means comparable to the estimates from the example with CLIN as the outcome (Table 1, bottom). Main/internal validation study-based analysis assuming the correct differential misclassification model produces reliable estimates of all four β coefficients, and excellent confidence interval (CI) coverage. In contrast, the main/internal analysis based on erroneously assuming non-differentiality produces average parameter estimates remarkably similar to the estimates reported in the lower half of Table 2. These are invalid except for the estimate of β_1, corresponding to the predictor (X_1) that was unassociated with sensitivity and specificity in model (14).

Simulation II: Misclassification of Case-Control Status

Table 5 summarizes simulations assessing the internal validation study-based methods under "case-control" sampling as previously described. The version of model (1) for generating data was as follows:

$$\text{logit}\,[\Pr(Y=1)] = \beta_0 + \beta_1 X_1 + \beta_2 X_2,$$

where X_1 is standard normally distributed and X_2 is a Bernoulli(0.5) binary predictor. The true regression coefficients were $(\beta_0,\beta_1,\beta_2) = (-0.4, 2.0, 0.5)$. For both scenarios in Table 5, approximately 5,200 observations were first generated via the above model in a cross-sectional manner. Error-prone (Y*) values were then generated, potentially allowing sensitivity and specificity to vary with X_2 [i.e., assuming $SE_t = \Pr(Y^*=1 \mid Y=1, X_2=t)$ and $SP_t = \Pr(Y^*=0 \mid Y=0, X_2=t)$ (t=0,1)]. To mimic case-control sampling, we utilized 100% of data records with Y*=1 in each case but retained only a 5% random sample of those with Y*=0. Under these conditions, each simulated "case-control" sample contained approximately 1,500 observations, of which 500 were randomly selected into an internal validation sample. The main/internal validation study likelihood to analyze each data set is specified in Eq. (7)–(8).

The top half of Table 5 summarizes results for a non-differential case, in which $SE_1=SP_1=SE_0=SP_0=0.8$. As noted above under "Comments Regarding Case-Control Data," maximum likelihood estimates of the SE and SP parameters differed from the true value of 0.8 on average, reflecting the "operating" sensitivity and specificity under "case" oversampling. However, estimates of β_1 and β_2 are quite reliable, with means near the true values of 2 and 0.5 and near-nominal CI coverage. The bottom half of Table 5 summarizes a differential case, where $SE_1=0.8$, $SP_1=0.7$, $SE_0=0.6$, and $SP_0=0.9$. Note that the coefficient (β_1), corresponding to the predictor (X_1) that was not associated with sensitivity and specificity, remains validly estimated. As also mentioned above, however, validity for estimating β_2 is lost subsequent to X_2's direct association with sensitivity and specificity. In both cases shown in Table 5, "naïve" analysis based on Y* for case-control status yielded severe bias.

DISCUSSION

We have considered the problem of outcome misclassification in logistic regression, with emphasis on clearly specifying likelihood functions corresponding to main/external and main/internal validation study designs. This emphasis distinguishes our work from related prior references in the epidemiologic literature,[18, 21–22] which do not pursue the incorporation of validation data. Although validation data-based maximum likelihood methods are outlined in the comprehensive text of Carroll et al.,[11] the treatment there is purposefully general and therefore made without a real-data example or facilitating computations. With the practicing epidemiologist in mind, we have sought to motivate such methods for handling outcome misclassification with a real-world study, and to make them fully accessible via user-friendly programs that directly reflect the likelihood specifications and utilize common software for optimization.[31]

Our treatment includes detailed evaluation of maximum likelihood methodology via simulation studies. These simulations, along with the HIV study example, clearly demonstrate the importance of internal validation subsampling when misclassification is differential. The results in Tables 2 and 4 illustrate that outcome misclassification adjustment via an erroneous assumption of non-differentiality may offer only marginal improvement over "naïve" analysis based on the error-prone outcome (Y*).

We have demonstrated how the methods considered here can be used in the case-control setting, for which little discussion about outcome misclassification in logistic regression appears in the literature and for which prior proposals[18] were not applicable. Assuming appropriate model specifications, we find that the maximum likelihood approach for the main/internal validation design illustrated here remains directly applicable in case-control studies with random "case" (i.e., based on Y*) oversampling under non-differential outcome misclassification. While further investigation of the impact of outcome-dependent sampling is warranted when misclassification is differential with respect to covariates, empirical studies suggest that the maximum likelihood approach maintains validity for estimating primary regression parameters associated with predictor variables that are not associated with sensitivity and specificity values.

Future work could involve extensions of past research on cost-efficiency[39] to the logistic regression setting considered here, because the ultimate appeal of main/internal validation study designs is their potential for conserving resources. Somewhat along these lines, we experimented with further simulations under the same conditions as were assumed in producing Table 4, but varying the size of the internal validation subsample. We found that decreasing the validation sampling fraction to select as few as 5% of the 1,000 subjects very seldom produced numerical problems with the maximum likelihood routine, despite expected increases in variability of the adjusted log odds ratio estimates. From a practical standpoint, the simulation program used to produce the results in Table 4 is a sharable resource that could aid an investigator in determining the validation fraction indicated for a particular study, and provide insight into the cost-efficiency of a main/internal validation design.

There may also be interest in regression-based methods to adjust for outcome misclassification in situations where no gold standard exists, but one has access to replicates of an error-prone outcome measure or to a diagnostic measure viewable as an "alloyed" gold standard.[40–41] Additionally, there would be value in making methods[29] that non-parametrically estimate the distribution of $Y^* \mid (Y, \mathbf{X})$ more readily accessible in practice. Nevertheless, the logistic regression approach advocated in Eq. (9)–(10) facilitates likelihood ratio testing and is potentially robust when all predictors in Eq. (9) are categorical, given the freedom to saturate that model.

Acknowledgments

The following is a list of the HIV EPIDEMIOLOGY RESEARCH STUDY GROUP: Robert S. Klein, Ellie Schoenbaum, Julia Arnsten, Robert D. Burk, Chee Jen Chang, Penelope Demas, and Andrea Howard, from Montefiore Medical Center and the Albert Einstein College of Medicine; Paula Schuman, and Jack Sobel, from the Wayne State University School of Medicine; Anne Rompalo, David Vlahov, and David Celentano, from the Johns Hopkins University School of Medicine; Charles Carpenter, and Kenneth Mayer, from the Brown University School of Medicine; Ann Duerr, Lytt I. Gardner, Charles M. Heilig, Scott Holmberg, Denise Jamieson, Jan Moore, Ruby Phelps, Dawn Smith, and Dora Warren, from the Centers for Disease Control and Prevention; and Katherine Davenny, from the National Institute of Drug Abuse.

Funding: This work was supported by National Institute of Nursing Research Grant 1RC4NR012527-01, by National Institute of Environmental Health Sciences Grant 2R01-ES012458-5, and by PHS Grant UL1 RR025008 from the Clinical and Translational Science Award Program, National Institutes of Health, National Center for Research Resources. The HER Study was supported by the Centers for Disease Control and Prevention: U64/CCU106795, U64/CCU206798, U64/CCU306802, and U64/CCU506831.

REFERENCES

1. Thomas D, Stram D, Dwyer J. Exposure measurement error: influence on exposure-disease relationships and methods of correction. Ann Rev Public Health. 1993; 14:69–93. [PubMed: 8323607]

2. Bross IDJ. Misclassification in 2×2 tables. Biometrics. 1954; 10:478–486.
3. Barron BA. The effects of misclassification on the estimation of relative risk. Biometrics. 1977; 33:414–418. [PubMed: 884199]
4. Kleinbaum, D.; Kupper, L.; Morgenstern, H. Epidemiologic Research: Principles and Quantitative Methods. Lifetime Learning; Belmont, CA: 1982.
5. Greenland S, Kleinbaum DG. Correcting for misclassification in two-way tables and matched-pair studies. Int J Epidemiol. 1983; 12:93–97. [PubMed: 6840961]
6. Greenland S. Variance estimation for epidemiologic effect estimates under misclassification. Stat Med. 1988; 7:745–757. [PubMed: 3043623]
7. Marshall RJ. Validation study methods for estimating proportions and odds ratios with misclassified data. J Clin Epidemiol. 1990; 43:941–947. [PubMed: 2213082]
8. Morrissey MJ, Spiegelman D. Matrix methods for estimating odds ratios with misclassified exposure data: extensions and comparisons. Biometrics. 1999; 55:338–344. [PubMed: 11318185]
9. Lyles RH. A note on estimating crude odds ratios in case-control studies with differentially misclassified exposure. Biometrics. 2002; 58:1034–1037. [PubMed: 12495160]
10. Greenland S. Maximum-likelihood and closed-form estimators of epidemiologic measures under misclassification. J Stat Plan Inf. 2008; 138:528–538.
11. Carroll, RJ.; Ruppert, D.; Stefanski, LA.; Crainiceanu, CM. Measurement Error in Nonlinear Models. Second Edition. Chapman and Hall; London: 2006.
12. Lyles RH, Zhang F, Drews-Botsch C. Combining internal and external validation data to correct for exposure misclassification: a case study. Epidemiol. 2007; 18:321–328.
13. Rosner B, Spiegelman D, Willett WC. Correction of logistic regression relative risk estimates and confidence intervals for measurement error: the case of multiple covariates measured with error. Am J Epidemiol. 1990; 132:734–745. [PubMed: 2403114]
14. Rosner B, Spiegelman D, Willett WC. Correction of logistic regression relative risk estimates and confidence intervals for random within-subject measurement error. Am J Epidemiol. 1992; 136:1400–1413. [PubMed: 1488967]
15. Spiegelman D, Casella M. Fully parametric and semi-parametric regression models for common events with covariate measurement error in main study/validation study designs. Biometrics. 1997; 53:395–400. [PubMed: 9192443]
16. Spiegelman D, Carroll RJ, Kipnis V. Efficient regression calibration for logistic regression in main study/internal validation study designs with an imperfect reference instrument. Stat Med. 2001; 20:139–160. [PubMed: 11135353]
17. Thurston SW, Williams PL, Hauser R, Hu H, Hernandez-Avila M, Spiegelman D. A comparison of regression calibration approaches for designs with internal validation data. J Stat Plan Inf. 2003; 131:175–190.
18. Magder LS, Hughes JP. Logistic regression when the outcome is measured with uncertainty. Am J Epidemiol. 1997; 146:195–203. [PubMed: 9230782]
19. Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. J Roy Stat Soc B. 1977; 39:1–38.
20. Neuhaus JM. Bias and efficiency loss due to misclassified responses in logistic regression. Biometrika. 1999; 86:843–855.
21. Lash TL, Fink AK. Semi-automated sensitivity analysis to assess systematic errors in observational data. Epidemiol. 2003; 14:451–458.
22. Fox MP, Lash TL, Greenland S. A method to automate probabilistic sensitivity analyses of misclassified binary variables. Int J Epidemiol. 2005; 34:1370–1376. [PubMed: 16172102]
23. Stamey JD, Young DA, Seaman JW. A Bayesian approach to adjust for diagnostic misclassification between two mortality causes in Poisson regression. Stat Med. 2008; 27:2440–2452. [PubMed: 17979218]
24. Meier AS, Richardson BA, Hughes JP. Discrete proportional hazards models for mismeasured outcomes. Biometrics. 2003; 59:947–954. [PubMed: 14969473]
25. Margaret AS. Incorporating validation subsets into discrete proportional hazards models for mismeasured outcomes. Stat Med. 2008; 27:5456–5470. [PubMed: 18613225]

26. Paulino CD, Soares P, Neuhaus J. Binomial regression with misclassification. Biometrics. 2003; 59:670–675. [PubMed: 14601768]
27. McInturff P, Johnson WO, Cowling D, Gardner IA. Modelling risk when binary outcomes are subject to error. Stat Med. 2004; 23:1095–1109. [PubMed: 15057880]
28. Gerlach R, Stamey J. Bayesian model selection for logistic regression with misclassified outcomes. Stat Modelling. 2007; 7:255–273.
29. Pepe MS. Inference using surrogate outcome data and a validation sample. Biometrika. 1992; 79:355–365.
30. Cheng KF, Hsueh HM. Correcting bias due to misclassification in the estimation of logistic regression models. Stat Prob Letters. 1999; 44:229–240.
31. SAS Institute, Inc.. SAS/STAT 9.1 User's Guide. SAS Institute, Inc.; Cary, NC: 2004.
32. Prentice RL, Pyke R. Logistic disease incidence models and case-control studies. Biometrika. 1979; 66:403–411.
33. Carroll RJ, Wang S, Wang CY. Prospective analysis of logistic case-control studies. J Am Stat Assoc. 1995; 90:157–169.
34. Smith DK, Warren DL, Vlahov D, Schuman P, Stein MD, Greenberg BL, et al. 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. Am J Epidemiol. 1997; 146:459–469. [PubMed: 9290506]
35. Amsel R, Totten PA, Spiegel CA, Chen KC, Eschenbach D, Holmes KK. Nonspecific vaginitis: diagnostic criteria and microbial and epidemiologic associations. Am J Med. 1983; 74:14–22. [PubMed: 6600371]
36. Nugent RP, Krohn MA, Hillier SL. Reliability of diagnosing bacterial vaginosis is improved by a standardized method of gram stain interpretation. J Clin Microbiol. 1991; 29:297–301. [PubMed: 1706728]
37. Jamieson DJ, Duerr A, Klein RS, Paramsothy P, Brown W, Cu-Uvin S, Rompalo A, Sobel J. Longitudinal analysis of bacterial vaginosis: findings from the HIV epidemiology research study. Obstet Gyn. 2001; 98:656–663.
38. Gallo MF, Jamieson DJ, Cu-Uvin S, Rompalo A, Klein RS, Sobel JD. Accuracy of clinical diagnosis of bacterial vaginosis by human immunodeficiency virus infection status. Sex Transm Dis. in press.
39. Spiegelman D, Gray R. Cost-efficient study designs for binary response data with Gaussian covariate measurement error. Biometrics. 1991; 47:851–869. [PubMed: 1789885]
40. Wacholder S, Armstrong B, Hartge P. Validation studies using an alloyed gold standard. Am J Epidemiol. 1993; 137:1251–1258. [PubMed: 8322765]
41. Brenner H. Correcting for exposure misclassification using an alloyed gold standard. Epidemiol. 1996; 7:406–410.

Table 1

Logistic regression results on 924 women at their 4th study visit

Outcome Variable: LAB status (gold standard)		
Variable	**β (std. error)**	**Estimated OR (95% CI)**
Race/Ethnicity (Black vs. White/Hispanic)	0.95 (0.15)	2.59 (1.94 – 3.46)
Risk Group (IDU vs. sex)	0.40 (0.14)	1.50 (1.13 – 1.98)
HIV Status (positive vs negative)	0.17 (0.15)	1.19 (0.89 – 1.60)
Age (years)	−0.06 (0.01)	0.95 (0.92 – 0.97)

Outcome Variable: CLIN status (error-prone)		
Race/Ethnicity (Black vs. White/Hispanic)	0.82 (0.17)	2.28 (1.63 – 3.20)
Risk Group (IDU vs. sex)	0.98 (0.17)	2.68 (1.93 – 3.73)
HIV Status (positive vs negative)	−0.35 (0.17)	0.71 (0.51 – 0.99)
Age (years)	−0.07 (0.01)	0.94 (0.91 – 0.96)

IDU indicates intravenous drug use.

Table 2

Results of maximum likelihood analysis of main / internal validation study data on 924 women ($n_m = 624$; $n_v = 300$) at their 4th study visit: Estimates of primary model parameters

Assuming differential misclassification[a]			
Variable	**$\hat{\beta}$ (std. error)**	**Estimated OR (95% CI)**	
Race/Ethnicity (Black vs. White/Hispanic)	1.13 (0.22)	3.10 (2.01 – 4.77)	
Risk Group (IDU vs. sex)	0.62 (0.24)	1.86 (1.17 – 2.95)	
HIV Status (positive vs negative)	0.20 (0.25)	1.22 (0.75 – 1.98)	
Age (years)	−0.07 (0.02)	0.94 (0.90 – 0.97)	

Assuming non-differential misclassification[b]			
Variable	**$\hat{\beta}$ (std. error)**	**Estimated OR (95% CI)**	
Race/Ethnicity (Black vs. White/Hispanic)	1.16 (0.22)	3.18 (2.07 – 4.86)	
Risk Group (IDU vs. sex)	0.98 (0.21)	2.67 (1.76 – 4.04)	
HIV Status (positive vs negative)	−0.10 (0.22)	0.90 (0.59 – 1.38)	
Age (years)	−0.08 (0.02)	0.93 (0.89 – 0.96)	< 0.001

[a] Sensitivity and specificity assumed to vary with the binary variables HIV risk group, HIV status, and Age (> median vs. < median), via model (14)

[b] No covariates affecting sensitivity and specificity; this assumption is not supported by the data ($P<0.001$)

Table 3

Results of maximum likelihood analysis of main / internal validation study data on 924women ($n_m = 624$; $n_v = 300$) at their 4th study visit: Estimates of secondary model parameters

Estimates based on logistic model in model (14)[a]				
Intercept $\hat{\theta}_0$	LAB BV $\hat{\theta}_1$	HIV risk group $\hat{\theta}_2$	HIV status $\hat{\theta}_3$	Age (> vs. < median) $\hat{\theta}_4$
−2.36 (0.33)	2.58 (0.36)	0.81 (0.23)	−0.60 (0.24)	−0.43 (0.22)

SE and SP estimates for subgroups based on model (14)				
HIV risk group	HIV status	Age	$\hat{SE}$	$\hat{SP}$
sex	negative	< median	0.55 (0.070)	0.91 (0.026)
sex	negative	> median	0.45 (0.071)	0.94 (0.018)
sex	positive	< median	0.40 (0.052)	0.95 (0.016)
sex	positive	> median	0.31 (0.051)	0.97 (0.011)
IDU	negative	< median	0.74 (0.059)	0.82 (0.048)
IDU	negative	> median	0.65 (0.064)	0.88 (0.033)
IDU	positive	< median	0.60 (0.053)	0.90 (0.031)
IDU	positive	> median	0.50 (0.053)	0.93 (0.021)

SE indicates sensitivity; SP, specificity.

[a] Standard errors. For sensitivity and specificity, standard errors are based on the multivariate delta method

Table 4

Results of simulations designed to mimic conditions of illustrative example based on study data[a]

Model	$\hat{\beta}_1$ (SD) [95% CI coverage]	$\hat{\beta}_2$ (SD) [95% CI coverage]	$\hat{\beta}_3$ (SD) [95% CI coverage]	$\hat{\beta}_4$ (SD) [95% CI coverage]
"Naïve"	0.62 (0.16) [13.4%]	0.96 (0.17) [45.0%]	−0.35 (0.17) [9.8%]	−0.06 (0.01) [81.0%]
Main/internal validation[b]	1.15 (0.23) [95.3%]	0.63 (0.25) [94.7%]	0.22 (0.25) [95.6%]	−0.07 (0.02) [95.6%]
Main/internal validation[c]	1.12 (0.22) [95.3%]	1.02 (0.21) [50.8%]	−0.10 (0.22) [69.3%]	−0.08 (0.02) [89.1%]

[a] 1000 simulations; 300 internal validation and 700 main study observations per simulation Main/internal validation study likelihood defined in Eq. (11)–(12) True model (1) parameters: $(\beta_0,\beta_1,\beta_2,\beta_3,\beta_4) = (0.85, 1.13, 0.62, 0.20, -0.07)$; θ and SE/SP parameters set equal to estimates given in Table 3

[b] Based on correct modeling of differential SE/SP parameters via model (14)

[c] Based on incorrect non-differentiality assumption

Table 5

Results of simulations assessing main/internal validation study-based analysis under case-control sampling[a,b]

Non-differential Case: $SE_1 = SP_1 = SE_0 = SP_0 = 0.8$				
Model	$\hat{\beta}_1$ **(SD)**	**95% CI coverage**	$\hat{\beta}_2$ **(SD)**	**95% CI coverage**
"Naïve"	0.30 (0.06)	0.0%	0.07 (0.16)	22.7%
Main/internal validation[b]	2.04 (0.20)	95.3%	0.51 (0.27)	95.1%

Differential Case: $SE_1=0.8$, $SP_1=0.7$, $SE_0=0.6$, $SP_0=0.9$				
Model	$\hat{\beta}_1$ **(SD)**	**95% CI coverage**	$\hat{\beta}_2$ **(SD)**	**95% CI coverage**
"Naïve"	0.30 (0.07)	0.0%	1.28 (0.16)	0.0%
Main/internal validation[b]	2.02 (0.20)	95.2%	−0.07 (0.30)	48.0%

[a] 1000 simulations in each case; 500 internal validation and roughly 1,000 main study observations per simulation based on 100% and 5% sampling of cases and controls, respectively. True model (1) parameters: $(\beta_0,\beta_1\beta_2) = (-0.4, 2.0, 0.5)$

[b] Main/internal validation study likelihood defined in Eq. (7)–(8)