Published in final edited form as:

Epidemiology. 2013 March ; 24(2): 215–223. doi:10.1097/EDE.0b013e31827f4fa9.

Sensitivity Analysis for Nonignorable Missingness and Outcome Misclassification from Proxy Reports

Michelle Shardell^{1,*}, Eleanor Simonsick², Gregory E. Hicks³, Barbara Resnick⁴, Luigi Ferrucci², and Jay Magaziner¹

¹Department of Epidemiology and Public Health, University of Maryland School of Medicine

²National Institute on Aging

³Department of Physical Therapy, University of Delaware

⁴University of Maryland School of Nursing

Abstract

Researchers often recruit proxy respondents, such as relatives or caregivers, for epidemiologic studies of older adults when study participants are unable to provide self-reports (e.g., due to illness or cognitive impairment). In most studies involving proxy-reported outcomes, proxies are recruited only to report on behalf of participants who have missing self-reported outcomes; thus, either a proxy report or participant self-report, but not both, is available for each participant. When outcomes are binary and investigators conceptualize participant self-reports as gold standard measures, substituting proxy reports in place of missing participant self-reports in statistical analysis can introduce misclassification error and lead to biased parameter estimates. However, excluding observations from participants with missing self-reported outcomes may also lead to bias. We propose a pattern-mixture model that uses error-prone proxy reports to reduce selection bias from missing outcomes, and we describe a sensitivity analysis to address bias from differential outcome misclassification. We perform model estimation with high-dimensional (e.g., continuous) covariates using propensity-score stratification and multiple imputation. We apply the methods to the Second Cohort of the Baltimore Hip Studies, a study of elderly hip-fracture patients, to assess the relation between type of surgical treatment and perceived physical recovery. Simulation studies show that the proposed methods perform well. We provide SAS programs in the eAppendix to enhance the methods' accessibility.

Older adults, especially those who are sick or cognitively impaired, may be unwilling or unable to respond to interview questions,^{1,2} potentially leading to nonignorable missing outcomes and, in most cases, inaccurate conclusions.^{3–5} To address this problem, investigators often recruit proxy respondents, such as relatives or caregivers, to report outcomes on behalf of nonresponding participants.⁶ When the objective is measuring older adults' subjective perceptions, such as depressive symptoms or perceived ability to perform activities of daily living, gerontology researchers regard participant self-reports as the gold standard and proxy reports as error-prone alternatives. However, substituting proxy reports for missing participant self-reports may lead to misclassification error and biased estimation.^{7,8} The bias is sometimes large enough to reverse study conclusions.⁹ An additional challenge in most studies using proxies is that a response for each participant is

*Corresponding Author: Michelle Shardell, University of Maryland School of Medicine, 660 West Redwood Street, Baltimore, Maryland 21201, mshardel@epi.umaryland.edu, phone: 410-706-8563, fax: 410-706-3808.

SDC Supplemental digital content is available through direct URL citations in the HTML and PDF versions of this article (www.epidem.com). This content is not peer-reviewed or copy-edited; it is the sole responsibility of the author.

obtained from either the participant or a proxy, but not both. Thus, proxy reports cannot be empirically validated.

Bias from outcome misclassification has been well studied.^{10–16} However, current statistical methods for misclassification assume that gold-standard outcomes are missing by design, where researchers purposely choose instead to collect more accessible error-prone outcomes for all participants. In contrast, we consider studies where gold-standard outcomes are missing by happenstance by a possibly nonignorable mechanism, and researchers collect error-prone outcomes only for participants with missing gold-standard outcomes. The left third of Table 1 exemplifies an ideal study without outcome misclassification whereas the middle third represents the usual data structure of so-called “main-study” designs with outcome misclassification that can be analyzed using existing statistical methods. The right third of Table 1 exemplifies the typical data structure of main-study designs in gerontology research with proxy respondents. In this study design, some participants have only a gold-standard self-report, the remaining participants have only an error-prone proxy report, and the missingness mechanism for participant self-reports is unknown to investigators and unestimable from observed data. Studies with the highest percentage of nonresponse rely most heavily on proxy respondents and are therefore the most susceptible to misclassification bias. Such studies often focus on particularly vulnerable populations, including persons with stroke,¹⁷ cancer,¹⁸ or osteoporotic fractures.¹⁹

Statistical methods for sensitivity analysis have been developed separately to address selection bias^{20–22} and misclassification bias;^{23–25} however, sensitivity-analysis methods that jointly address both sources of bias are crucially needed. We aim to provide statistical methods that address both outcome misclassification and missingness for epidemiologists working with proxy-reported outcomes. To achieve our aim, we use missing-data methods to address participants’ nonresponse while leveraging proxy reports. Specifically, we use pattern-mixture models^{26,27} to encode assumptions about missingness by relating an unestimable function of missing participant self-reports to an estimable function of observed proxy reports. The proposed methods for binary outcomes extend those developed by Shardell and colleagues⁷ for measurement error of proxy-reported continuous outcomes. We use pattern-mixture models to incorporate proxy responses into analysis, and we adapt propensity-score methods to estimate covariate-standardized proportions. We apply these methods to the Second Cohort of the Baltimore Hip Studies, an epidemiologic study of elderly hip fracture patients.²⁸

METHODS

We aim to assess the relation between a binary (0, 1) exposure X and true binary (0, 1) outcome Y standardized for (potentially vector-valued) covariates Z . We denote the desired probability as $p_x = \sum_z P(Y = 1 / X = x, Z = z) \times P(Z = z)$, the x -specific probability that $Y = 1$ standardized for Z . If Y were fully observed without error, we could estimate p_x using multiple methods. If Z is low dimensional, we could estimate proportions of Y within X - and Z -specific strata and average over a standard Z distribution.²⁹ If Z is high dimensional, we could perform propensity-score stratification. This approach entails estimating exposure propensity scores,^{30,31} $PS(z) = P(X = 1 / Z = z)$, for example by logistic regression, creating a small number of strata, estimating outcome proportions within substrata defined by X and propensity-score strata, and averaging over the propensity-score stratum proportions; essentially treating propensity-score stratum as a low-dimensional covariate.³¹ Another approach to estimating p_x using high-dimensional covariates is computing the reciprocal of the probability of observed exposure, $W(z) = X/PS(z) + (1 - X)/[1 - PS(z)]$, then using inverse probability weighting to directly calculate standardized odds ratios, risk ratios, or risk differences.³² For example, SAS (Cary, North Carolina) PROC GENMOD³³ using an

identity link with $W(z)$ entered into a WEIGHT statement estimates $p_1 - p_0$ using the study sample as the standard population.

We consider studies where, for some participants, instead of observing Y (participant self-report), we only observe a binary (0, 1) error-prone proxy report, Y^* . Gerontology researchers commonly substitute Y^* for missing Y to create a completed dataset and use statistical methods described in the previous paragraph. This naïve “participant + proxy” approach generally produces biased results,¹⁰ where the degree of bias depends on discordance between Y^* and Y (misclassification bias). Other researchers exclude observations with missing Y . This “participant-only” analysis also may produce biased results,³⁴ where the degree of bias depends on the missingness mechanism (selection bias) and analytic method. Recent work uses a structural framework to describe missingness mechanisms whereby participant-only (i.e., complete-case) analysis leads to bias.^{4,5}

Researchers also sometimes ignore proxy reports and use conventional, though principled, methods to handle missing outcomes. This approach involves positing explicit assumptions about the missing-data mechanism that cannot be verified from the data, and that require additional notation. Let R denote the observation status of Y : $R = 1$ if Y is observed, and $R = 0$ if Y is missing (Y^* observed). Assumptions about missingness are specified by noting that the joint distribution of Y and R , $P(Y = y, R = r / X = x, Z = z)$, can be factored as either $P(R = r / X = x, Z = z, Y = y) \times P(Y = y / X = x, Z = z)$ or $P(Y = y / X = x, Z = z, R = r) \times P(R = r / X = x, Z = z)$ for $y = 0, 1$; $r = 0, 1$. The first factorization, called a selection model, is the basis for inverse-probability of response methods.³⁴ The second factorization, called a pattern-mixture model,^{26,27} has received less attention in the epidemiology literature than selection models (although an accessible introduction exists²¹).

Selection models directly estimate $P(Y = y / X = x, Z = z)$, the quantity needed to calculate p_x ; however, encoding the missingness mechanism requires researchers to specify how Y affects the probability of being observed (or selected), an abstract concept and difficult task. In contrast, pattern-mixture models encode more transparent assumptions about missingness, where researchers specify how the distribution of Y differs between the two values, or “patterns”, of R . The disadvantage of pattern-mixture models is an additional calculation needed to obtain $P(Y = y / X = x, Z = z)$. Namely, $P(Y = y / X = x, Z = z)$ for $y = 0, 1$ is computed as a weighted average, or “mixture,” of two probabilities, one for participants with $R = 0$ and another for participants with $R = 1$:

$$\begin{aligned} P(Y=y|X=x, Z=z) \\ = P(Y=y|X=x, Z=z, R=0) \times P(R=0|X=x, Z=z) \\ + P(Y=y|X=x, Z=z, R=1) \times P(R=1|X=x, Z=z). \end{aligned} \quad (1)$$

$P(R = r / X = x, Z = z)$ ($r = 0, 1$) and $P(Y = y / X = x, Z = z, R = 1)$ are estimable from the data; estimation of $P(Y = y / X = x, Z = z, R = 0)$ requires additional assumptions about the missingness mechanism.

Researchers often assume that outcomes are “missing at random;”³ that is, Y and R are independent given X and Z . This assumption is encoded in selection models by setting $P(R = r / X = x, Z = z, Y = y) = P(R = r / X = x, Z = z)$ and in pattern-mixture models by setting $P(Y = y / X = x, Z = z, R = 0) = P(Y = y / X = x, Z = z, R = 1)$. The assumed relation between estimable and unestimable quantities in pattern-mixture models is called a “pattern-mixture restriction.”^{26,27} When outcomes are missing at random, researchers can easily perform unbiased estimation of $P(Y = y / X = x, Z = z)$ using likelihood-based participant-only methods. When outcome data are nonignorably missing, or missing not at random, Y

and R remain dependent given X and Z , and researchers must specify exactly how Y relates to R . However, this relation is difficult to intuit. In contrast, researchers may more easily intuit the relation between Y and Y^* . In some subject-matter areas, including gerontology, intuition is enhanced by published validation studies reporting concordance between Y and Y^* in similar study populations. Y^* can be particularly helpful because it is a purposeful attempt to measure missing Y , and therefore is plausibly more strongly predictive of Y than any other measured variable. Researchers can most directly use Y^* by fitting pattern-mixture models relating $P(Y = y / X = x, Z = z, R = 0)$ to $P(Y^* = y^* / X = x, Z = z, R = 0)$ rather than to $P(Y = y / X = x, Z = z, R = 1)$. Participant + proxy analysis is a special case, implicitly encoding the assumption $P(Y = y / X = x, Z = z, R = 0) = P(Y^* = y^* / X = x, Z = z, R = 0)$.

Leveraging Proxy Reports for Sensitivity Analysis

We propose a pattern-mixture restriction using proxy reports to inform the distribution of missing participant self-reports, analogous to a strategy used in recent work on continuous outcomes.⁷ In the present case of binary outcomes, we can adapt ideas from the misclassification literature. We first note that $P(Y^* = y^* / X = x, Z = z, R = 0)$ is estimable and can be written^{11,16} as

$$P(Y^* = y^* | X = x, Z = z, R = 0) = \sum_{y=0}^1 P(Y^* = y^* | Y = y, X = x, Z = z, R = 0) \times P(Y = y | X = x, Z = z, R = 0). \quad (2)$$

The first term in the summation is a function of sensitivity or specificity for $y = 1$ or 0 , respectively. Let $Sens(x, z) = P(Y^* = 1 / Y = 1, X = x, Z = z, R = 0)$ and $Spec(x, z) = P(Y^* = 0 / Y = 0, X = x, Z = z, R = 0)$ denote sensitivity and specificity, respectively, among participants with $R = 0$. We estimate $P(Y = y / X = x, Z = z, R = 0)$ by imposing a pattern-mixture restriction that relates unestimable $Sens(x, z)$ and $Spec(x, z)$ to $P(Y^* = y^* / X = x, Z = z, R = 0)$. The advantage of this pattern-mixture restriction is that sensitivity and specificity are well-known, easily interpretable concepts. Researchers perform a sensitivity analysis by comparing results for a range of supplied values for $Sens(x, z)$ and $Spec(x, z)$. To ensure that estimated $P(Y = y / X = x, Z = z, R = 0)$ lies between 0 and 1, $Sens(x, z)$ and $Spec(x, z)$ must satisfy the following constraints:

$$1 - Spec(x, z) < P(Y^* = 1 | X = x, Z = z, R = 0) < Sens(x, z). \quad (3)$$

To avoid positing separate values of sensitivity and specificity for each x and z , investigators often assume non-differential misclassification²³ (independence of x and z):

$$Sens(x, z) = P(Y^* = 1 | Y = 1, R = 0) \quad \text{and} \quad Spec(x, z) = P(Y^* = 0 | Y = 0, R = 0).$$

An advantage of assuming non-differential misclassification is simplicity, because researchers need only a single value for sensitivity and specificity, but the assumption is often unrealistic. Furthermore, assuming non-differential misclassification can violate Eq. (3), requiring ad hoc solutions.²³ To overcome these limitations while retaining the advantage of simplicity, we use “exponential tilt models”³⁵ to relate $P(Y^* = y^* / Y = y, X = x, Z = z, R = 0)$ to $P(Y^* = y^* / X = x, Z = z, R = 0)$. Exponential tilt models are commonly used in the missing-data literature to model pattern-mixture restrictions^{20,21,36,37} and are helpful here because they allow sensitivity and specificity to vary by x and z using a small

number of user-specified “tuning parameters” while satisfying Eq. (3). The general form for an exponential tilt model in this context is:

$$P(Y^*=y^*|Y=y, X=x, Z=z, R=0) = \frac{P(Y^*=y^*|X=x, Z=z, R=0)e^{q(y,y^*)}}{P(Y^*=0|X=x, Z=z, R=0)e^{q(y,0)} + P(Y^*=1|X=x, Z=z, R=0)e^{q(y,1)}}. \quad (4)$$

The investigator uses the unestimable user-specified function $e^{q(y,y^*)}$ to encode assumptions about sensitivity and specificity. The assumed $e^{q(y,y^*)}$ helps to estimate $P(Y^*=y^*|Y=y, X=x, Z=z, R=0)$ by “tilting” $P(Y^*=y^*|X=x, Z=z, R=0)$ toward higher or lower values depending on y . By convention, we set $q(1, 0) = q(0, 1) = 0$. We denote $q_{sens} \equiv q(1, 1)$ and $q_{spec} \equiv q(0, 0)$, which are the tuning parameters that affect sensitivity and specificity, respectively. Specifying $q_{sens} > 0$ and $q_{spec} > 0$ satisfies the constraints Eq. (3) for all x and z .

To understand how to interpret q_{sens} and q_{spec} and how they affect sensitivity and specificity, consider an example where $P(Y^*=1|X=x, Z=z, R=0) = 0.60$ for a particular x and z , and $q_{sens} = 1.75$ and $q_{spec} = 0.50$. Then

$$\begin{aligned} Sens(z, x) &= 0.60e^{q_{sens}} / [0.40e^0 + 0.60e^{q_{sens}}] = 0.60e^{1.75} / [0.40 + 0.60e^{1.75}] = 0.896, \\ Spec(z, x) &= 0.40e^{q_{spec}} / [0.40e^{q_{spec}} + 0.60e^0] = 0.40e^{0.50} / [0.40e^{0.50} + 0.60] = 0.524. \end{aligned}$$

In words, q_{sens} encodes the assumption that conditioning on $Y=1$ increases the probability that $Y^*=1$ from 0.60 to 0.896 given $X=x, Z=z$, and $R=0$. Also, $e^{q_{sens}}$ is the factor by which the odds of $Y^*=1$ is assumed to increase after conditioning on $Y=1$, given $X=x, Z=z$, and $R=0$. The odds here increase from $0.6/0.4$ to $(0.6/0.4) \times e^{1.75}$. Similarly, q_{spec} encodes the assumption that conditioning on $Y=0$ increases the probability that $Y^*=0$ from 0.40 to 0.524 given $X=x, Z=z$, and $R=0$. Additionally, $e^{q_{spec}}$ is the factor by which the odds of $Y^*=0$ is assumed to increase after conditioning on $Y=0$, given $X=x, Z=z$, and $R=0$. In this example, the odds increase from $0.4/0.6$ to $(0.4/0.6) \times e^{0.50}$. Larger q_{sens} and q_{spec} encode higher sensitivity and specificity, respectively. Setting $q_{sens} = q_{spec} = \infty$ implies $Sens(x, z) = Spec(x, z) = 1$ and is equivalent to participant + proxy analysis. We provide guidance on selecting q_{sens} and q_{spec} in the eAppendix.

Once $Sens(x, z)$ and $Spec(x, z)$ are estimated using the pattern-mixture restrictions, $P(Y=y|X=x, Z=z, R=0)$ can be estimated from Eq. (2) using classic “matrix-method” type calculations described in the misclassification literature.¹³

$$P(Y=1|X=x, Z=z, R=0) = \frac{P(Y^*=1|X=x, Z=z, R=0) + Spec(x, z) - 1}{Sens(x, z) + Spec(x, z) - 1}. \quad (5)$$

Plugging estimated $P(Y=1|X=x, Z=z, R=0)$ into Eq. (1) estimates $P(Y=y|X=x, Z=z)$. While $P(Y=y|X=x, Z=z)$ is useful as an x - and z -specific predicted probability, it does not automatically produce a Z -adjusted measure of association because it is a weighted average of (presumably) non-linear functions of X and Z . For example, if $P(Y=y|X=x, Z=z, R=r)$ is linear-logistic, then $P(Y=y|X=x, Z=z)$ will not be linear-logistic and will not provide a Z -adjusted odds ratio. However, standardized probabilities p_x can be used to estimate Z -standardized odds ratios, risk ratios, or risk differences. If Z is low-dimensional, then $P(Y=y|X=x, Z=z)$ and standard errors can be directly estimated and used to calculate p_x and corresponding standard errors.²⁹ The eAppendix includes a SAS program for this approach using PROC NL MIXED.³³

We subsequently describe computationally accessible extensions of this approach for handling high-dimensional covariates using propensity score stratification³¹ and multiple imputation³⁸ followed by inverse probability-weighted estimation.³² Information on proxy characteristics, such as the proxy's relationship or living arrangement with the participant, can also be used to estimate p_x ; the eAppendix includes an exposition demonstrating how to incorporate auxiliary proxy characteristics into the analysis.

Propensity Score Stratification with High-Dimensional Covariates

Propensity scores are probabilities of an exposure level as a function of covariates. Propensity-score stratification creates a low-dimensional variable from high-dimensional covariates that can be used in the method described in the previous section. Five to 10 strata are often recommended (e.g., quintiles or deciles) for bias reduction.³⁹ Let S denote stratum and n_s denote number of strata. The pattern-mixture model in Eq. (1) is replaced by

$$\begin{aligned} P(Y=y|X=x, S=s) \\ = P(Y=y|X=x, S=s, R=0) \times P(R=0|X=x, S=s) \quad (6) \\ + P(Y=y|X=x, S=s, R=1) \times P(R=1|X=x, S=s), \end{aligned}$$

for $s = 1, \dots, n_s$ and $P(Y = y / X = x, S = s, R = 0)$ is found by solving Eq. (2) with $S = s$ replacing $Z = z$. Similarly, stratum-specific sensitivity $Sens(x, s) \equiv P(Y^* = 1 / Y = 1, X = x, S = s, R = 0)$ and specificity $Spec(x, s) \equiv P(Y^* = 0 / Y = 0, X = x, S = s, R = 0)$ are found using the exponential tilt models in Eq. (4) with $S = s$ replacing $Z = z$. When $q_{sens} > 0$ and $q_{spec} > 0$, analogous constraints in Eq. (3) are satisfied with S and s replacing Z and z , respectively. Computing Eq. (5) with S and s replacing Z and z , respectively, will estimate $P(Y = 1 / X = x, S = s, R = 0)$ between 0 and 1. Plugging $P(Y = 1 / X = x, S = s, R = 0)$ into Eq. (6) produces stratum-specific probabilities.

Computing the weighted average of the stratum-specific probabilities,

$p_x = \sum_{s=1}^{n_s} P(Y=y|X=x, S=s) \times P(S=s)$, completes the approach. We provide SAS³³ code for propensity-score stratification and estimation in the eAppendix.

Multiple Imputation with High-Dimensional Covariates

Multiple imputation generates M estimates (usually $M=10$ to 20) of missing Y .³⁸ Each set of imputed Y completes a data set to be analyzed using standard complete-data methods, resulting in M parameter estimates. The parameter estimates and corresponding variances are then combined to form final parameter estimates and variances.³⁸ Multiple imputation involves two general steps: first, estimating parameters of an imputation model (in this case, $P(Y = y / X = x, Z = z, R = 0)$), and then randomly sampling values of missing outcomes from the imputation model. Specific steps for imputation with proxy respondents are as follows:

1. Sample participants with $R = 0$ with replacement (bootstrap sample).
2. Estimate parameters of $P(Y^* = 1 / X = x, Z = z, R = 0)$ (e.g., by logistic regression) using the bootstrapped sample.
3. Calculate fitted values of $P(Y^* = 1 / X = x, Z = z, R = 0)$ using the original participants with $R = 0$.
4. Plug fitted values of $P(Y^* = 1 / X = x, Z = z, R = 0)$ into Eq. (5) to estimate $P(Y = 1 / X = x, Z = z, R = 0)$.
5. For each participant with $R = 0$, draw a value from $P(Y = 1 / X = x, Z = z, R = 0)$.

Repeating steps 1–5 M times produces M completed data sets. Once the data sets have been completed, researchers estimate p_x for each dataset (M estimates) using complete-data methods such as propensity-score stratification³¹ or inverse probability weighting³² and calculate the final estimate using the combining rules described by Rubin.³⁸ To calculate the M estimates of p_x from inverse probability weighting, first calculate the propensity scores $PS(z) = P(X = 1 / Z = z)$ from the original data (e.g., using logistic regression), then compute weights as $W(z) = X/PS(z) + (1 - X)/[1 - PS(z)]$. Lastly, for each completed dataset, regress Y on X using $W(z)$ as the weight. The eAppendix includes a SAS³³ macro for multiple imputation and analysis using inverse probability weighting.

SIMULATION STUDIES

We evaluated the finite-sample performances of both pattern-mixture model estimation procedures by simulating 1000 studies of 500 observations each. We set $q_{sens} = 1.75$ and $q_{spec} = 0.50$ and simulated three covariates $\mathbf{Z} = (Z_1, Z_2, Z_3)$: Z_1 was Bernoulli(0.5), and (Z_2, Z_3) were bivariate normal with mean $(0, 0.1z_1)$, variance $(1, 1)$ and correlation 0.5. X was Bernoulli with $\text{Logit}[P(X = 1 / \mathbf{Z} = \mathbf{z})] = -0.3 + 1.0z_1 + 0.2z_2 - 0.2z_3$; and R was Bernoulli with $\text{Logit}[P(R = 1 / X = x, \mathbf{Z} = \mathbf{z})] = 0.2 + 0.3z_1 - 0.4z_2 - 0.5z_3 + 0.6x$. When $R = 1$, Y was simulated from a Bernoulli distribution with $\text{Logit}[P(Y = 1 / X = x, \mathbf{Z} = \mathbf{z}, R = 1)] = 0.2 - 0.5z_1 - 0.05z_2 + 0.1z_3 + 1.0x$. When $R = 0$, Y^* was simulated from a Bernoulli distribution with $\text{Logit}[P(Y^* = 1 / X = x, \mathbf{Z} = \mathbf{z}, R = 0)] = 0.3 + 0.3z_1 - 0.1z_2 - 0.1z_3 + 1.0x$. The median (interquartile range) of sensitivity and specificity were 0.90 (0.88–0.92) and 0.50 (0.44–0.53), respectively. We performed propensity-score standardization using quintiles ($n_s = 5$) and multiple imputation using $M = 20$, both with assumed q_{sens} and q_{spec} at the correct values and incorrectly assuming $q_{sens} = q_{spec} = 99$, an arbitrarily large value, to demonstrate that participant + proxy analysis is equivalent to the pattern-mixture model assuming perfect sensitivity and specificity.

We compared the proposed methods with participant-only and participant + proxy analyses. We assessed performance using percent relative bias and by comparing estimated standard errors (SEs) with empirical standard errors (ESEs) to determine whether the methods accurately estimate standard errors. We calculated percent relative bias as $(\text{Estimate} - \text{True Value}) \times 100 / \text{True Value}$. SE was the average of 1000 estimated standard errors, and ESE was the standard deviation of 1000 parameter estimates. We also determined empirical coverage of 95% confidence intervals, $\text{Estimate} \pm 1.96 \times \text{SE}$.

Table 2 shows that participant-only and participant + proxy analyses produced biased estimates. In contrast, pattern-mixture models estimated at the correct q_{sens} and q_{spec} produced unbiased estimates. Propensity-score stratification produced accurate standard errors; multiple imputation produced conservatively large standard errors, resulting in coverage that was slightly higher than 95% when q_{sens} and q_{spec} were correctly specified. Also, pattern-mixture modeling incorrectly assuming $q_{sens} = q_{spec} = 99$ produced estimates within rounding error of those from participant + proxy analysis. Contour plots of propensity-score stratification and multiple imputation in the Figure 1 show that estimates were sensitive to q_{sens} and q_{spec} , particularly at values close to 0.

EXAMPLE: THE BALTIMORE HIP STUDIES

We applied propensity-score stratification and multiple imputation to the Second Cohort of the Baltimore Hip Studies, a study of predictors of recovery from hip fracture among older adults.²⁸ The current analysis assessed whether surgical treatment type (internal fixation versus arthroplasty) relates to perceived recovery of independent mobility, operationalized as self-reported ability to walk 10 feet without human assistance two months after fracture.

Analysis included 501 participants recruited from eight Baltimore area hospitals between 1 January 1990 and 15 June 1991. The University of Maryland's Institutional Review Board approved the study. Informed consent was obtained from participants or their proxies. Three-hundred eighteen participants underwent internal fixation, treatment with hardware such as metal plates and screws, and 183 underwent arthroplasty, partial or total hip replacement. Self-reported mobility was obtained from 362 participants, among whom 281 (78%) self-reported being independently mobile. Proxies were recruited for the remaining 139 participants; 63 (45%) reported that the participant was independently mobile.

We adjusted all analyses for sex (395 women, 106 men), years of age (mean=80.7 [SD=7.4]; range 65–104), and number of comorbid conditions (mean=3.2 [SD=2.1]; range 0–12). We performed participant-only analysis, participant + proxy analysis, and pattern-mixture models estimated using propensity score stratification and multiple imputation followed by inverse probability weighting.³² Pattern-mixture modeling was initially performed using two sets of tuning parameters (q_{sens} , q_{spec}): (1.75, 0.50) and (0.50, 1.75). The first set encodes the assumption that conditioning on $Y = 1$ increases the odds that $Y^* = 1$ by a factor of $e^{1.75}$ and conditioning on $Y = 0$ increases the odds that $Y^* = 0$ by a factor of $e^{0.50}$, leading to median (interquartile range) sensitivity and specificity of 0.81 (0.78–0.85) and 0.68 (0.62–0.72), respectively (proxies better at correctly identifying mobility independence than dependence); the second set encodes the assumption that conditioning on $Y = 1$ increases the odds that $Y^* = 1$ by a factor of $e^{0.50}$, and conditioning on $Y = 0$ increases the odds that $Y^* = 0$ by a factor of $e^{1.75}$, leading to median (interquartile range) sensitivity and specificity of 0.55 (0.51–0.62) and 0.88 (0.85–0.90), respectively (proxies better at correctly identifying mobility dependence than independence). We also obtained estimates setting $q_{sens} = q_{spec} = 99$.

Table 3 shows that after standardization for sex, age, and comorbidities, the estimated probability of independent mobility was lower for participants who underwent internal fixation compared with those who underwent arthroplasty, but the magnitude depended on analytic method. Participant-only analysis estimated the largest absolute difference (−0.10 [SE=0.04]); pattern-mixture modeling using both propensity-score stratification and multiple imputation with $q_{sens} = 1.75$ and $q_{spec} = 0.50$ estimated the smallest absolute difference (both methods: −0.05 [SE=0.04]). Estimates from participant + proxy analysis and pattern-mixture modeling with $q_{sens} = q_{spec} = 99$ were within a rounding error of each other.

DISCUSSION

This paper provides analytical approaches to address missing participant self-report outcomes and error-prone proxy reports. We focused on “main study” designs of cross-sectional and cohort studies for analyzing existing datasets with proxy responses. We were particularly motivated by aging research where proxy respondents are commonly used. Studies of older adults with proxies include the Women's Health and Aging Study,⁴⁰ Established Populations for Epidemiologic Studies of the Elderly,⁴¹ and the National Health and Nutrition Examination Survey.⁴² However, researchers can also apply the methods to studies of other vulnerable persons who need proxy reports, such as patient populations.

Our proposed methods overcome limitations of conventional approaches for binary outcomes with misclassification error by addressing the problem that participants with proxy reports are systematically different from participants with self-reports. However, we retain the benefits of conventional approaches^{10,11,23} by facilitating sensitivity analysis using a small number of interpretable user-specified tuning parameters. We have also provided methods for using auxiliary proxy data, guidance on performing sensitivity analysis, and SAS programs in the eAppendix for increased usability.

Principled statistical methods that use proxy responses are also helpful for epidemiologists who design studies and must decide whether to recruit proxies. Proxies are recruited because their responses are presumed highly predictive of participant responses. If this premise is true, and if researchers can posit the relation between proxy and participant responses (e.g., using published validation studies), then collecting proxy responses and using them in statistical analysis is worthwhile. However, if other completely-observed variables are more predictive of participant responses (and missingness) than proxy responses, and if researchers have a good understanding of the missingness mechanism (e.g., missing at random or missing not at random), then costs (time, money, etc.) of recruiting proxies may outweigh benefits, and conventional, principled statistical analysis with missing outcomes suffices.³⁴

Future work includes methods that accommodate internal validation data in the form of validation proxies who provide reports on behalf of participants without missing self-reports (i.e., “main study + internal validation data” designs).⁷ Researchers can use validation proxies to estimate sensitivity and specificity; although these estimates would be biased (at best, generalizable to participants with $R = 1$), the estimates would provide reasonable “anchors” for sensitivity analysis. Additionally, although the methods and accompanying SAS programs have been illustrated using logistic regression to estimate propensity scores and $P(Y^* = 1 / X = x, Z = z, R = 0)$, it is worthwhile to assess the performance of flexible non-parametric methods such as machine-learning techniques in this context.⁴³ Lastly, we conceptualized participant self-reports as gold standard measures here, because the Baltimore Hip Studies researchers were interested in patients’ perceptions of their mobility. If researchers consider self-report as an imperfect reference instrument (an “alloyed” gold standard)⁴⁴ or as one of many ratings of an underlying latent construct,⁸ then different methods may be required.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank Dawn Alley for providing thoughtful comments on an earlier draft of this manuscript.

Financial Support: This research was supported by NIH grants K25 AG034216 (Shardell), R01 AG06322 (Magaziner), R01 AG09902 (Magaziner)

References

1. Ferrucci L, Guralnik JM, Studenski S, Fried LP, Cutler GB Jr, Walston JD. Designing randomized, controlled trials aimed at preventing or delaying functional decline and disability in frail, older persons: a consensus report. *J Am Geriatr Soc.* 2004; 52:625–634. [PubMed: 15066083]
2. Mody L, Miller DK, McGloin JM, et al. Recruitment and retention of older adults in aging research. *J Am Geriatr Soc.* 2008; 56:2340–2348. [PubMed: 19093934]
3. Rubin DB. Inference and missing data. *Biometrika.* 1976; 63:581–592.
4. Daniel RM, Kenward MG, Cousens SN, De Stavola BL. Using causal diagrams to guide analysis in missing data problems. *Stat Methods Med Res.* 2012; 21:243–256. [PubMed: 21389091]
5. Westreich D. Berkson’s bias, selection bias, and missing data. *Epidemiology.* 2012; 23:159–164. [PubMed: 22081062]
6. Magaziner, J. The use of proxy respondents in health studies of the aged. In: Wallace, RB.; Woolson, RF., editors. *The Epidemiologic Study of the Elderly*. New York: Oxford University Press; 1992. p. 120-129.

7. Shardell M, Hicks GE, Miller RR, Langenberg P, Magaziner J. Pattern-mixture models for analyzing normal outcome data with proxy respondents. *Stat Med.* 2010; 29:1522–1538. [PubMed: 20535763]
8. Snow AL, Cook KF, Lin PS, Morgan RO, Magaziner J. Proxies and other external raters: methodological considerations. *Health Serv Res.* 2005; 40:1676–1693. [PubMed: 16179002]
9. Shardell M, Alley DE, Miller RR, Hicks GE, Magaziner J. Comparing reports from hip-fracture patients and their proxies: implications on evaluating sex differences in disability and depressive symptoms. *J Aging Health.* 2012; 24:367–383. [PubMed: 22210805]
10. Carroll, RJ.; Ruppert, D.; Stefanski, LA.; Crainiceanu, CM. *Measurement Error in Nonlinear Models: A Modern Perspective.* 2. London: Chapman and Hall; 2006.
11. Magder LS, Hughes JP. Logistic regression when the outcome is measured with uncertainty. *Am J Epidemiol.* 1997; 146:195–203. [PubMed: 9230782]
12. Bross IDJ. Misclassification in 2×2 tables. *Biometrics.* 1954; 10:478–486.
13. Barron BA. The effects of misclassification on the estimation of relative risk. *Biometrics.* 1977; 33:414–418. [PubMed: 884199]
14. Greenland S, Kleinbaum DG. Correcting for misclassification in two-way tables and matched-pair studies. *Int J Epidemiol.* 1983; 12:93–97. [PubMed: 6840961]
15. Neuhaus JM. Analysis of clustered and longitudinal binary data subject to response misclassification. *Biometrics.* 2002; 58:675–683. [PubMed: 12230004]
16. Neuhaus JM. Bias and efficiency loss due to misclassified responses in logistic regression. *Biometrika.* 1999; 86:843–855.
17. Oczkowski C, O'Donnell M. Reliability of proxy respondents for patients with stroke: a systematic review. *J Stroke Cerebrovasc Dis.* 2010; 19:410–416. [PubMed: 20554222]
18. Tang ST, McCorkle R. Use of family proxies in quality of life research for cancer patients at the end of life: a literature review. *Cancer Invest.* 2002; 20:1086–1104. [PubMed: 12449742]
19. Magaziner J, Zimmerman SI, Gruber-Baldini AL, Hebel JR, Fox KM. Proxy reporting in five areas of functional status. Comparison with self-reports and observations of performance. *Am J Epidemiol.* 1997; 146:418–428. [PubMed: 9290502]
20. Shardell M, El-Kamary SS. Sensitivity analysis of informatively coarsened data using pattern mixture models. *J Biopharm Stat.* 2009; 19:1018–1038. [PubMed: 20183462]
21. Shardell M, Scharfstein DO, Vlahov D, Galai N. Sensitivity analysis using elicited expert information for inference with coarsened data: Illustration of censored discrete event times in ALIVE. *Am J Epidemiol.* 2008; 168:1460–1469. [PubMed: 18952850]
22. Wood AM, White IR, Thompson SG. Are missing outcome data adequately handled? A review of published randomized controlled trials in major medical journals. *Clin Trials.* 2004; 1:368–376. [PubMed: 16279275]
23. Lyles RH, Lin J. Sensitivity analysis for misclassification in logistic regression via likelihood methods and predictive value weighting. *Stat Med.* 2010; 29:2297–2309. [PubMed: 20552681]
24. Lash TL, Fink AK. Semi-automated sensitivity analysis to assess systematic errors in observational data. *Epidemiology.* 2003; 14:451–458. [PubMed: 12843771]
25. 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]
26. Little RJ. Pattern-mixture models for multivariate incomplete data. *J Am Stat Assoc.* 1993; 88:125–134.
27. Little RJ, Wang Y. Pattern-mixture models for multivariate incomplete data with covariates. *Biometrics.* 1996; 52:98–111. [PubMed: 8934587]
28. Magaziner J, Hawkes W, Hebel JR, et al. Recovery from hip fracture in eight areas of function. *J Gerontol A Biol Sci Med Sci.* 2000; 55:M498–M507. [PubMed: 10995047]
29. Rothman, KJ.; Greenland, S.; Lash, TL. *Modern Epidemiology.* 3. Philadelphia: Wolters Kluwer Health/Lippincott Williams and Wilkins; 2008.
30. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika.* 1983; 70:41–55.

31. Rosenbaum PR, Rubin DB. Reducing bias in observational studies using subclassification on the propensity score. *J Am Stat Assoc.* 1984; 79:516–524.
32. Sato T, Matsuyama Y. Marginal structural models as a tool for standardization. *Epidemiology.* 2003; 14:680–686. [PubMed: 14569183]
33. SAS Institute, Inc. SAS/STAT 9.2 User's Guide. Cary, NC: SAS Institute, Inc; 2009.
34. Little, RJA.; Rubin, DB. *Statistical Analysis with Missing Data.* 2. New York: Wiley; 2002.
35. Barndorff-Nielsen, OE.; Cox, DR. *Asymptotic Techniques for use in Statistics.* London: Chapman and Hall; 1989.
36. Wang C, Daniels MJ, Scharfstein DO, Land S. A Bayesian shrinkage model for incomplete longitudinal binary data with application to the Breast Cancer Prevention Trial. *J Am Stat Assoc.* 2010; 105:1333–1346. [PubMed: 21516191]
37. Birmingham J, Rotnitzky A, Fitzmaurice GM. Pattern-mixture and selection models for analysing longitudinal data with monotone missing patterns. *J Royal Stat Soc Ser B.* 2003; 65:275–297.
38. Rubin, DB. *Multiple Imputation for Nonresponse in Surveys.* New York: Wiley; 2004.
39. D'Agostino RB. Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Stat Med.* 1998; 17:2265–2281. [PubMed: 9802183]
40. Kasper JD, Shapiro S, Guralnik JM, Bandeen-Roche KJ, Fried LP. Designing a community study of moderately to severely disabled older women: the Women's Health and Aging Study. *Ann Epidemiol.* 1999; 9:498–507. [PubMed: 10549883]
41. Cornoni-Huntley, J.; Blazer, DG.; Lafferty, ME.; Everett, DF.; Brock, DB.; Farmer, ME. *Established Populations for Epidemiologic Studies of the Elderly.* Bethesda, Md: National Institute on Aging; 1990.
42. National Center for Health Statistics. [Accessed December 5, 2011] National Health and Nutrition Examination Survey Documentation. Available at: <http://www.cdc.gov/nchs/nhanes/nhanes.htm>
43. Lee BK, Lessler J, Stuart EA. Improving propensity score weighting using machine learning. *Stat Med.* 2010; 29:337–346. [PubMed: 19960510]
44. Brenner H. Correcting for exposure misclassification using an alloyed gold standard. *Epidemiology.* 1996; 7:406–410. [PubMed: 8793367]

Figure 1.

Figure Simulation study results. $\hat{p}_0$ from A) propensity-score stratification and B) multiple imputation; $\hat{p}_1$ from C) propensity-score stratification and D) multiple imputation; $\hat{p}_1 - \hat{p}_0$ from E) propensity-score stratification and F) multiple imputation. True values denoted by*.

Table 1

Main Study Data Structures (Y = Gold Standard Outcome, Y^* = Error-Prone Version, ? = Missing).

Ideal Study		Typical Study		Study with Proxies	
ID	Y	Y^*	ID	Y	Y^*
1	Y_1	?	1	?	Y^*_1
2	Y_2	?	2	?	Y^*_2
3	Y_3	?	3	?	Y^*_3
4	Y_4	?	4	?	Y^*_4
5	Y_5	?	5	?	Y^*_5
6	Y_6	?	6	?	Y^*_6
7	Y_7	?	7	?	Y^*_7
8	Y_8	?	8	?	Y^*_8

Table 2

Simulation Study Results.^a

Parameter	Analytic Method	Estimate	% Relative Bias	SE	ESE	Coverage ^b
p_1	Participant Only ^c	0.716	21	0.033	0.033	4.2
	Participant + Proxy ^d	0.690	17	0.029	0.028	6.6
	PMM-PS, Correct ^e	0.590	<1	0.027	0.026	95.1
	PMM-PS, Incorrect ^f	0.689	17	0.029	0.028	6.9
	PMM-MI, Correct ^g	0.591	<1	0.028	0.026	96.2
p_0	PMM-MI, Incorrect ^h	0.690	17	0.031	0.028	9.7
	Participant Only	0.484	20	0.046	0.046	56.3
	Participant + Proxy	0.528	31	0.035	0.034	4.2
	PMM-PS, Correct	0.401	<1	0.028	0.028	94.4
	PMM-PS, Incorrect	0.529	31	0.034	0.024	3.9
$p_1 - p_0$	PMM-MI, Correct	0.401	<1	0.031	0.028	96.5
	PMM-MI, Incorrect	0.529	31	0.038	0.035	7.6
	Participant Only	0.232	23	0.057	0.057	86.7
	Participant + Proxy	0.161	-14	0.045	0.046	90.7
	PMM-PS, Correct	0.188	<1	0.039	0.039	95.1
	PMM-PS, Incorrect	0.161	-14	0.045	0.046	90.6
	PMM-MI, Correct	0.189	<1	0.042	0.039	95.8
	PMM-MI, Incorrect	0.161	-14	0.049	0.046	93.6

^a1000 simulations, 500 observations. True parameters: $p_0 = 0.402$, $p_1 = 0.590$, $p_1 - p_0 = 0.188$.^bPercent of 95% confidence intervals covering true parameter values;^cParticipant Only = Analysis excluding observations with $R = 0$;^dParticipant + Proxy = Analysis substituting missing Y with Y^* ;^ePMM-PS, Correct = Pattern-mixture model estimated using propensity-score stratification correctly assuming $q_{sens} = 1.75$ and $q_{spec} = 0.50$;^fPMM-PS, Incorrect = PMM-PS incorrectly assuming $q_{sens} = q_{spec} = 99$;

$\hat{g}_{\text{PMM-MI}}$, Correct = Pattern-mixture model estimated using multiple imputation and inverse probability weighting correctly assuming $q_{\text{sens}} = 1.75$ and $q_{\text{spec}} = 0.50$;

$\hat{h}_{\text{PMM-MI}}$, Incorrect = PMM-MI incorrectly assuming $q_{\text{sens}} = 99$ and $q_{\text{spec}} = 99$.

Table 3
Results for 501 Patients in the Second Cohort of the Baltimore Hip Studies at Two Months Post-Fracture.

Analytic Method	q_{sens}	q_{spec}	Proportion Mobile ^a							
			Internal Fixation			Arthroplasty			Difference	
			$\hat{p}_{IF}$	SE	$\hat{p}_A$	SE	$\hat{p}_{IF} - \hat{p}_A$	SE		
Participant Only ^b			0.74	0.03	0.84	0.03	-0.10	0.04		
Participant + Proxy ^c			0.66	0.03	0.73	0.03	-0.07	0.04		
PMM-PS ^d	1.75	0.50	0.61	0.02	0.66	0.03	-0.05	0.04		
	0.50	1.75	0.73	0.02	0.81	0.02	-0.08	0.03		
	99	99	0.66	0.02	0.73	0.03	-0.07	0.04		
PMM-MI ^e	1.75	0.50	0.61	0.02	0.66	0.03	-0.05	0.04		
	0.50	1.75	0.74	0.03	0.81	0.02	-0.07	0.04		
	99	99	0.67	0.03	0.73	0.04	-0.06	0.05		

^a Able to walk 10 feet without human assistance two months after hip fracture.

^b Participant Only = Analysis excluding observations with missing participant self reports;

^c Participant + Proxy = Analysis substituting missing participant self reports with proxy reports;

^d PMM-PS = Pattern-mixture model estimated using propensity-score stratification;

^e PMM-MI = Pattern-mixture model estimated using multiple imputation and inverse probability weighting.