Published in final edited form as:

Stat Med. 2014 November 10; 33(25): 4437–4452. doi:10.1002/sim.6238.

Statistical analysis with missing exposure data measured by proxy respondents: a misclassification problem within a missing-data problem

Michelle Shardell^{1,*} and Gregory E. Hicks²

¹Department of Epidemiology and Public Health, University of Maryland, Baltimore, Maryland, USA

²Department of Physical Therapy, University of Delaware, 303 McKinly Lab Newark, Delaware 19716, USA

Abstract

In studies of older adults, researchers often recruit proxy respondents, such as relatives or caregivers, when study participants cannot provide self-reports (e.g., due to illness). Proxies are usually only sought to report on behalf of participants with missing self-reports; thus, either a participant self-report or proxy report, but not both, is available for each participant. Furthermore, the missing-data mechanism for participant self-reports is not identifiable and may be nonignorable. When exposures are binary and participant self-reports are conceptualized as the gold standard, substituting error-prone proxy reports for missing participant self-reports may produce biased estimates of outcome means. Researchers can handle this data structure by treating the problem as one of misclassification within the stratum of participants with missing self-reports. Most methods for addressing exposure misclassification require validation data, replicate data, or an assumption of nondifferential misclassification; other methods may result in an exposure misclassification model that is incompatible with the analysis model. We propose a model that makes none of the aforementioned requirements and still preserves model compatibility. Two user-specified tuning parameters encode the exposure misclassification model. Two proposed approaches estimate outcome means standardized for (potentially) high-dimensional covariates using multiple imputation followed by propensity-score methods. The first method is parametric and uses maximum likelihood to estimate the exposure misclassification model (i.e., the imputation model) and the propensity score model (i.e., the analysis model); the second method is non-parametric and uses boosted classification and regression trees to estimate both models. We apply both methods to a study of elderly hip-fracture patients.

Keywords

Exposure Misclassification; Gerontology; Missing Data; Proxy Respondents

*Correspondence to: Department of Epidemiology and Public Health, University of Maryland, Baltimore, Maryland 21201-1596, USA. mshardel@epi.umaryland.edu, Phone: 410-706-8563, Fax: 410-706-3808.

1. Introduction

Studies of older adults often include interview questions on participants' perceptions, attitudes, and beliefs. This information is used to measure disability, depressive symptoms, and other subjective constructs as predictors of physical decline and mortality. A common problem in assessment is that frail participants are often unable to respond to interview questions due to cognitive impairment or illness. Researchers are concerned about excluding this segment of the population from analysis, because doing so would lead to selection bias from missing data and would produce findings that represent only healthier members of the older adult population.

To prevent selection bias, researchers often attempt to recruit a proxy respondent to report on behalf of a participant who has missing data [1]. However, the proxy report is an error-prone version of the missing participant self-report. When the response options result in a binary exposure variable, the proxy report is likely subject to differential misclassification [2], and substituting the proxy report in place of the missing participant self-report in analysis can lead to misclassification bias either toward or away from the null. Furthermore, since most studies only recruit proxies to report on behalf of participants with missing data, the proxy reports cannot be empirically validated.

Statistical analysis with exposure misclassification is challenging, because the misclassification mechanism is not identified by the data and most methods require either available validation data, replicate data, or an assumption of nondifferential misclassification (see [3, 4] and references therein). Furthermore, most conventional methods for misclassification were developed to address exposures missing by design, not those missing by happenstance via an unidentifiable and potentially nonignorable mechanism. To overcome these limitations and address the data structure in studies with proxy respondents, we propose a statistical approach that treats the issue as a misclassification problem embedded within a missing-data framework. The proposed approach applies multiple imputation [5] to estimate missing exposures using proxy reports and applies propensity-score methods to calculate standardized outcome means. The exposure misclassification model is indexed by two user-specified tuning parameters that represent an assumed level of agreement between the observed proxy and missing participant responses. A sensitivity analysis can be performed by varying the tuning parameters.

The proposed method builds on our previous work addressing misclassified or mismeasured outcome data from proxy respondents [6, 7]. An additional complication of misclassified exposure data is the need to specify an exposure misclassification model (an imputation model), and ensure that it is compatible with the analysis model, which in this case is the propensity-score model [3, 4]. To handle this complication, we propose both parametric and non-parametric approaches to estimate both the imputation and analysis models, where parametric estimation is carried out using maximum likelihood and non-parametric estimation is carried out using boosted classification and regression trees (CART) [8–10].

We assess the methods' performance using a simulation study and apply the methods to a study of hip fracture patients to determine whether perceived rapid recovery of mobility is associated with two-year survival [11]. Throughout the paper, we conceptualize participant self-reports as the gold standard and proxy reports as an error-prone version to be consistent with the study design most commonly used in aging research [12].

2. Data and Model

Let Y denote an outcome variable and let X denote the true value of a binary exposure variable. Let $\mathbf{Z}$ denote a vector of covariates. Our goal is to estimate

$$\mu_x = \sum_{\mathbf{z}} E(Y|X=x, \mathbf{Z}=\mathbf{z}) \times f(\mathbf{Z}=\mathbf{z}), \quad (1)$$

the x -specific ($x = 0, 1$) mean of Y standardized for $\mathbf{Z}$. If $\mathbf{Z}$ comprises all confounders of the X - Y relation, then $\mu_1 - \mu_0$ equals the causal contrast $E[Y(1) - Y(0)]$, where $Y(x)$ denotes the potential outcome if exposure X were set to value x . Equation (1) is a special case of the g-formula proposed by Robins [13]. If X were completely observed without error, Equation (1) can be estimated using propensity-score methods [14–16], including inverse-probability weighting [17]. This relation can be seen by re-writing Equation (1) as

$$\mu_x = \sum_{\mathbf{z}} E(Y, X=x, \mathbf{Z}=\mathbf{z}) / P(X=x|\mathbf{Z}=\mathbf{z}), \quad (2)$$

where $P(X=x|\mathbf{Z}=\mathbf{z})$ is the propensity score of having $X=x$ given covariates $\mathbf{Z}=\mathbf{z}$. Consider a study with n participants where X_i , Y_i , and $\mathbf{Z}_i$ denote the variables for the i th participant, $i = 1, \dots, n$. A complete-data estimator for μ_x is essentially a Monte Carlo integration over the distribution of $\mathbf{Z}$ among those with $X=x$:

$$\frac{1}{n} \sum_{i=1}^n I(X_i=x) Y_i / \hat{P}(X_i=x|\mathbf{Z}_i), \quad (3)$$

where $I(\cdot)$ is the indicator function and $\hat{P}(X_i=x|\mathbf{Z}_i)$ is an estimator for $P(X_i=x|\mathbf{Z}_i)$. Thus, consistent estimation of μ_x can be accomplished by consistent estimation of $P(X=x|\mathbf{Z}=\mathbf{z})$. Equation (3) is the inverse-probability weighted estimator; the link between standardization and inverse-probability weighting is further described by Sato and Matsuyama [18].

Analysis is more complex when X is missing for some participants and is measured by proxy reports. In this scenario, unbiased estimation of $P(X=x|\mathbf{Z}=\mathbf{z})$ in Equation (2) using proxy reports requires an exposure model that 1) differentiates between data from participant self-reports and proxy reports, and 2) addresses the misclassification mechanism of proxy reports. Section 2.1 below differentiates participant self reports and proxy reports using a missing data framework, and Section 2.2 proposes a model for misclassification for the subgroup of participants who require proxy reports.

2.1. Missing Data Framework

We introduce additional notation to differentiate between participant self-reports and proxy reports and to express an exposure model using the missing-data framework. Let X^* denote the proxy-reported binary exposure variable. Let R be an indicator for participant response, where $R = 1$ when X is observed, and $R = 0$ when X is missing (X^* is observed).

Gerontology researchers may be concerned that X is missing not at random (MNAR) in the sense of Rubin [19], because older adults who are frailer and sicker may be more reliant on proxy respondents than their healthier counterparts. The implications of incomplete exposure data are somewhat counterintuitive. For example, if exposure is MNAR, but missingness does not depend on Y , then means of Y estimated using complete-case methods will be unbiased. In contrast, if exposure is missing at random (MAR), but missingness depends on Y , then means of Y estimated using complete-case methods will be biased [20, 21]. To minimize assumptions and maximize generality, we will presume that the exposure is MNAR, and that missingness may depend on Y (and covariates $\mathbf{Z}$).

When X is MNAR, then

$$P(X=x, R=r|\mathbf{Z}, R) \neq P(X=x|\mathbf{Z}, Y) \times P(R=r|\mathbf{Z}, Y),$$

for $x = (0, 1)$ and $r = (0, 1)$. In this case, researchers must specify the conditional joint distribution of X and R . However, since the joint distribution is not identifiable from the data, its specification depends on investigator-supplied assumptions about missingness. Researchers encode their assumptions by factoring the joint distribution into either a selection model,

$$P(X=x, R=r|\mathbf{Z}, Y) = P(R=r|X=x, \mathbf{Z}, Y) \times P(X=x|\mathbf{Z}, Y)$$

or a pattern-mixture model [22, 23],

$$P(X=x, R=r|\mathbf{Z}, Y) = P(X=x|\mathbf{Z}, Y, R=r) \times P(R=r|\mathbf{Z}, Y).$$

The selection model is often used, because it leads to direct estimation of $P(X=x|\mathbf{Z}, Y)$, where integration over Y is all that is needed to obtain propensity scores $P(X=x|\mathbf{Z})$ to plug into Equation (2). However, we will specify a pattern-mixture model, because it can easily accommodate proxy reports, X^* . To see this, note that $P(X=x|\mathbf{Z}, Y)$ can be calculated from a pattern-mixture model by

$$P(X=x|\mathbf{Z}, Y) = P(X=x|\mathbf{Z}, Y, R=0) \times P(R=0|\mathbf{Z}, Y) + P(X=x|\mathbf{Z}, Y, R=1) \times P(R=1|\mathbf{Z}, Y). \quad (4)$$

Thus, an unbiased estimate of $P(X=x|\mathbf{Z}, Y)$ can be obtained by plugging unbiased estimates into the right side of Equation (4). The quantity $P(X=x|\mathbf{Z}, Y, R=1)$ is estimable from observed participant self-reports; $P(R=r|\mathbf{Z}, Y)$, where $r = (0, 1)$, is also estimable. In

contrast, $P(X = x | \mathbf{Z}, Y, R = 0)$ is not estimable; however, $P(X^* = x^* | \mathbf{Z}, Y, R = 0)$, where $x^* = (0, 1)$, is estimable. Therefore, researchers can posit assumptions about $P(X = x | \mathbf{Z}, Y, R = 0)$ by relating it to $P(X^* = x^* | \mathbf{Z}, Y, R = 0)$.

Relating $P(X = x | \mathbf{Z}, Y, R = 0)$ to $P(X^* = x^* | \mathbf{Z}, Y, R = 0)$ in a pattern-mixture model is analogous to previously proposed models for handling proxy-reported outcome data [6, 7], but differs from conventional usage of pattern-mixture models. Conventionally, $P(X = x | \mathbf{Z}, Y, R = 0)$ would be specified by relating it to $P(X = x | \mathbf{Z}, Y, R = 1)$. For example, setting $P(X = 1 | \mathbf{Z}, Y, R = 0) = P(X = 1 | \mathbf{Z}, Y, R = 1)$ encodes the MAR assumption [19]. Analogously, a pattern-mixture model setting $P(X = 1 | \mathbf{Z}, Y, R = 0) = P(X^* = 1 | \mathbf{Z}, Y, R = 0)$ encodes the assumption that analysis using $RX + (1 - R)X^*$ as the exposure is unbiased.

Although researchers can ignore X^* and perform a principled analysis using one of many techniques [24], collecting and analyzing X^* may be useful because 1) specifying the relation between X and X^* may be more intuitive for subject-matter experts than specifying the relation between X and R due to the many published participant-proxy concordance studies available (see [1, 25] and references therein), and 2) X^* may be the strongest predictor of X available in the data set given that measuring X^* is an attempt to recover X .

2.2. Model for Misclassification within Stratum $R = 0$

Expressing the exposure model as a pattern-mixture model in Equation (4) shows that within the stratum $R = 0$, the problem simplifies to one of misclassification. Namely, estimating Equation (4) using X^* requires positing a relation between $P(X = 1 | \mathbf{Z}, Y, R = 0)$ and $P(X^* = 1 | \mathbf{Z}, Y, R = 0)$. To tackle this problem we adapt and apply previously published methods from misclassification error [3]. First, we express $P(X^* = 1 | \mathbf{Z}, Y, R = 0)$ as

$$P(X^* = 1 | \mathbf{Z}, Y, R = 0) = \sum_{x=0}^1 P(X^* = 1 | X = x, \mathbf{Z}, Y, R = 0) \times P(X = x | \mathbf{Z}, Y, R = 0), \quad (5)$$

and note that the first term in the summation can be interpreted as sensitivity or $1 -$ specificity for $x = 1$ or 0 , respectively [26]. The concepts of sensitivity and specificity are often used in the misclassification literature as interpretable, standardized measures of the assumed relation between X and X^* . Let $Sens(\mathbf{Z}, Y) = P(X^* = 1 | X = 1, \mathbf{Z}, Y, R = 0)$ and $Spec(\mathbf{Z}, Y) = P(X^* = 0 | X = 0, \mathbf{Z}, Y, R = 0)$ denote sensitivity and specificity, respectively, conditioned on $\mathbf{Z}$ and Y in the stratum $R = 0$. Therefore, a sensitivity analysis on Equation (4) boils down to varying the values of $Sens(\mathbf{Z}, Y)$ and $Spec(\mathbf{Z}, Y)$ in Equation (5).

Two related complications exist when positing values of $Sens(\mathbf{Z}, Y)$ and $Spec(\mathbf{Z}, Y)$. First, $Sens(\mathbf{Z}, Y)$ and $Spec(\mathbf{Z}, Y)$ must satisfy the constraint

$$1 - Spec(\mathbf{Z}, Y) < P(X^* = 1 | \mathbf{Z}, Y, R = 0) < Sens(\mathbf{Z}, Y) \quad (6)$$

to ensure that estimated values of $P(X = 1 | \mathbf{Z}, Y, R = 0)$ lie between 0 and 1. Second, sensitivity and specificity as (potentially high-dimensional) functions of $\mathbf{Z}$ and Y must be

posited. To simplify this problem, researchers often assume that misclassification is nondifferential, which implies that $Sens(\mathbf{Z}, Y)$ and $Spec(\mathbf{Z}, Y)$ are constant in $\mathbf{Z}$ and Y :

$$Sens(\mathbf{Z}, Y) \equiv P(X^* = 1 | X = 1, \mathbf{Z}, Y, R = 0) = P(X^* = 1 | X = 1, R = 0),$$

and

$$Spec(\mathbf{Z}, Y) \equiv P(X^* = 0 | X = 0, \mathbf{Z}, Y, R = 0) = P(X^* = 0 | X = 0, R = 0).$$

The benefit of assuming nondifferential classification is that only two unidentified user-specified values are needed: one each for sensitivity and specificity. However, the drawbacks of assuming nondifferential misclassification are two-fold. First, nondifferential misclassification is often implausible, particularly in the context of aging research with proxy respondents [2]. Second, assuming nondifferential misclassification may violate the constraints in Equation (6) for some values of $\mathbf{Z}$ and Y . To overcome these limitations while maintaining the advantage of requiring only two user-specified values, we instead encode assumptions about sensitivity and specificity using two tuning parameters and $P(X^* = 1 | \mathbf{Z}, Y, R = 0)$, where the latter quantity can be empirically estimated. Let q_{sens} and q_{spec} denote two unidentifiable tuning parameters that encode assumptions about sensitivity and specificity, respectively, and let

$$Sens(\mathbf{Z}, Y) \equiv \frac{P(X^* = 1 | \mathbf{Z}, Y, R = 0)e^{q_{sens}}}{P(X^* = 0 | \mathbf{Z}, Y, R = 0) + P(X^* = 1 | \mathbf{Z}, Y, R = 0)e^{q_{sens}}}, \quad (7)$$

and

$$Spec(\mathbf{Z}, Y) \equiv \frac{P(X^* = 0 | \mathbf{Z}, Y, R = 0)e^{q_{spec}}}{P(X^* = 0 | \mathbf{Z}, Y, R = 0)e^{q_{spec}} + P(X^* = 1 | \mathbf{Z}, Y, R = 0)}. \quad (8)$$

Using Equations (7) and (8), Equation (6) is satisfied by assuming $q_{sens}, q_{spec} > 0$. Equations (7) and (8) are examples of ‘exponential tilt models,’ [27] a common approach to modeling nonignorably missing data and performing sensitivity analysis (see [28, 29]).

The quantity $e^{q_{sens}}$ is interpreted as the odds ratio of $X^* = 1$ comparing participants with $X = 1$ to the whole population conditioned on $\mathbf{Z}$, Y and $R = 0$. Similarly, $e^{q_{spec}}$ is the odds ratio of $X^* = 0$ comparing participants with $X = 0$ to the whole population conditioned on $\mathbf{Z}$, Y and $R = 0$. For example, if $P(X^* = 1 | \mathbf{Z} = \mathbf{z}, Y = y, R = 0) = 0.75$ for a particular $(\mathbf{z}, y)$, and we assume that $e^{q_{sens}} = 3$ and $e^{q_{spec}} = 7$, then

$$\begin{aligned} Sens(\mathbf{Z} = \mathbf{z}, Y = y) &= 0.75e^{q_{sens}} / [0.25 + 0.75e^{q_{sens}}] = 0.75 \times 3 / [0.25 + 0.75 \times 3] = 0.90, \\ Spec(\mathbf{Z} = \mathbf{z}, Y = y) &= 0.25e^{q_{spec}} / [0.25e^{q_{spec}} + 0.75] = 0.25 \times 7 / [0.25 \times 7 + 0.75] = 0.70. \end{aligned}$$

In this example, setting $e^{q_{sens}} = 3$ implies that the odds of $X^* = 1$ increase from $0.75/0.25$ to $(0.75/0.25) \times 3$ after conditioning on $X = 1$; setting $e^{q_{spec}} = 7$ implies that the odds of $X^* = 0$ increase from $0.25/0.75$ to $(0.25/0.75) \times 7$ after conditioning on $X = 0$. Presumed sensitivity and specificity increase with increasing q_{sens} and q_{spec} , respectively. Assuming $q_{sens} = q_{spec} = \infty$ implies $Sens(\mathbf{Z}, Y) = Spec(\mathbf{Z}, Y) = 1$ and is equivalent to performing a complete-case analysis using $RX + (1 - R)X^*$ as the exposure.

Once $e^{q_{sens}}$ and $e^{q_{spec}}$ are specified, Equation (5) can be used to solve for $P(X = 1 | \mathbf{Z}, Y, R = 0)$. The solution is obtained by using the standard “matrix-method” calculations shown in the misclassification literature [30]:

$$P(X=1|\mathbf{Z}, Y, R=0) = \frac{P(X^*=1|\mathbf{Z}, Y, R=0) + Spec(\mathbf{Z}, Y) - 1}{Sens(\mathbf{Z}, Y) + Spec(\mathbf{Z}, Y) - 1}. \quad (9)$$

The quantity $P(X = 1 | \mathbf{Z}, Y, R = 0)$ is then used to calculate positive predictive value (PPV) and negative predictive value (NPV):

$$PPV(\mathbf{Z}, Y) \equiv P(X=1|X^*=1, \mathbf{Z}, Y, R=0) = Sens(\mathbf{Z}, Y) \frac{P(X=1|\mathbf{Z}, Y, R=0)}{P(X^*=1|\mathbf{Z}, Y, R=0)}, \quad (10)$$

$$NPV(\mathbf{Z}, Y) \equiv P(X=0|X^*=0, \mathbf{Z}, Y, R=0) = Spec(\mathbf{Z}, Y) \frac{P(X=0|\mathbf{Z}, Y, R=0)}{P(X^*=0|\mathbf{Z}, Y, R=0)}. \quad (11)$$

The quantities $PPV(\mathbf{Z}, Y)$ and $NPV(\mathbf{Z}, Y)$ form the basis of the exposure-misclassification imputation model. Also, plugging $P(X = 1 | \mathbf{Z}, Y, R = 0)$ into Equation (4) calculates $P(X = 1 | \mathbf{Z}, Y)$, which will be used to compute propensity scores and hence forms the basis of the analysis model.

In the next section, we describe how to estimate $PPV(\mathbf{Z}, Y)$, $NPV(\mathbf{Z}, Y)$, $P(X = 1 | \mathbf{Z}, Y)$, propensity scores, and standardized means μ_x while ensuring compatibility between the imputation models ($PPV(\mathbf{Z}, Y)$, $NPV(\mathbf{Z}, Y)$) and the analysis model ($P(X = 1 | \mathbf{Z})$).

3. Estimation

We propose a general strategy of using the exposure misclassification model to multiply impute missing X using $PPV(\mathbf{Z}, Y)$ and $NPV(\mathbf{Z}, Y)$. We then use the multiply imputed data to calculate standardized means μ_x via propensity-score methods, where the propensity score model is the analysis model. A challenge of this approach is to specify models for propensity scores, $PS(\mathbf{Z}) = P(X = 1 | \mathbf{Z})$, that are compatible with the imputation models, that is, so that there exists a full-data distribution that is consistent with all specified models. To achieve this goal, we first consider specification and estimation of parametric models that preserve model compatibility. We then describe estimation using non-parametric machine-learning methods that can avoid model incompatibility and other sources of model misspecification by reducing the number of needed models.

Consider a study with n participants where X_i , X_i^* , R_i , Y_i , and Z_i denote the variables for the i th participant, $i = 1, \dots, n$. Without loss of generality, let the respective distributions of X_i , X_i^* , R_i , and Y_i be denoted as

$$\pi_{X_i|Z_i Y_i 1} = P(X_i=1|Z_i, Y_i, R_i=1) \quad (12)$$

$$\pi_{X_i^*|Z_i Y_i 0} = P(X_i^*=1|Z_i, Y_i, R_i=0) \quad (13)$$

$$\pi_{R_i|Z_i Y_i} = P(R_i=1|Z_i, Y_i) \quad (14)$$

$$f_{Y_i|Z_i}(y_i) = f(Y_i=y_i|Z_i), \quad (15)$$

where $f(\cdot | \cdot)$ denotes a probability mass function for categorical outcomes and a probability density function for continuous outcomes. We henceforth may suppress the subscript i in notation if doing so does not cause ambiguity.

3.1. Parametric Estimation

Let the respective distributions for X_i , X_i^* , R_i , and Y_i be indexed by a finite number of parameters, $\beta = \{\beta_X, \beta_{X^*}, \beta_R, \beta_Y\}$, and be denoted as $\pi_{X_i|Z_i Y_i 1}(\beta_X)$, $\pi_{X_i^*|Z_i Y_i 0}(\beta_{X^*})$, $\pi_{R_i|Z_i Y_i}(\beta_R)$, and $f_{Y_i|Z_i}(y_i; \beta_Y)$. We estimate β , denoted $\hat{\beta} = \{\hat{\beta}_X, \hat{\beta}_{X^*}, \hat{\beta}_R, \hat{\beta}_Y\}$, by maximizing the observed-data likelihood $L(\beta)$ where

$$L(\beta) = \prod_{i=1}^n \{ [\pi_{X_i|Z_i Y_i 1}(\beta_X)^{X_i} (1 - \pi_{X_i|Z_i Y_i 1}(\beta_X))^{1-X_i} \pi_{R_i|Z_i Y_i}(\beta_R)]^{R_i} \times [\pi_{X_i^*|Z_i Y_i 0}(\beta_{X^*})^{X_i^*} (1 - \pi_{X_i^*|Z_i Y_i 0}(\beta_{X^*}))^{1-X_i^*} (1 - \pi_{R_i|Z_i Y_i}(\beta_R))^{1-R_i} \times f_{Y_i|Z_i}(y_i; \beta_Y) \}. \quad (16)$$

Equation (16) can be maximized by performing four separate regressions: 1) Regressing X_i on Z_i and Y_i using data from participants with $R_i = 1$ produces $\hat{\beta}_X$, which can be used to calculate fitted values $\hat{\pi}_{X_i|Z_i Y_i 1}(\hat{\beta}_X)$; 2) regressing X_i^* on Z_i and Y_i using data from participants with $R_i = 0$ produces $\hat{\beta}_{X^*}$, which can be used to calculate fitted values $\hat{\pi}_{X_i^*|Z_i Y_i 0}(\hat{\beta}_{X^*})$; 3) regressing R_i on Z_i and Y_i using data from all participants produces $\hat{\beta}_R$, which can be used to calculate fitted values $\hat{\pi}_{R_i|Z_i Y_i}(\hat{\beta}_R)$; and 4) regressing Y_i on Z_i using data from all participants produces $\hat{\beta}_Y$, which can be used to calculate estimated densities $\hat{f}_{Y_i|Z_i}(y_i; \hat{\beta}_Y)$. These parameter estimates can be used to calculate propensity scores, $PS(Z)$. If Y is categorical, then fitted values from maximizing $L(\beta)$ calculate

$$\widehat{PS}(Z_i; \hat{\beta}) = \sum_y \hat{P}(X_i=1|Z_i, Y_i=y; \hat{\beta}_X, \hat{\beta}_{X^*}, \hat{\beta}_R) \hat{f}_{Y_i}(y; \hat{\beta}_Y), \quad (17)$$

where, by plugging estimates into Equation (4),

$$\hat{P}(X_i=1|\mathbf{Z}_i, Y_i; \hat{\beta}_X, \hat{\beta}_{X^*}, \hat{\beta}_R) = \pi_{X_i|\mathbf{Z}_i Y_i 1}(\hat{\beta}_X) \times \hat{\pi}_{R_i|\mathbf{Z}_i Y_i}(\hat{\beta}_R) + \frac{\hat{\pi}_{X^*|\mathbf{Z}_i Y_i 0}(\hat{\beta}_{X^*}) + S\hat{p}ec(\mathbf{Z}, Y) - 1}{S\hat{e}ns(\mathbf{Z}, Y) + S\hat{p}ec(\mathbf{Z}, Y) - 1} \times (1 - \hat{\pi}_{R_i|\mathbf{Z}_i Y_i}(\hat{\beta}_R)),$$

and $S\hat{e}ns(\mathbf{Z}, Y)$ and $S\hat{p}ec(\mathbf{Z}, Y)$ are estimates of $Sens(\mathbf{Z}, Y)$ and $Spec(\mathbf{Z}, Y)$, respectively, found by plugging $\pi_{X^*|\mathbf{Z}_i Y_i 0}(\hat{\beta}_{X^*})$ into Equations (7) and (8). If Y is continuous, Monte Carlo integration can calculate $\widehat{PS}(\mathbf{Z}_i; \hat{\beta})$. The first step is to simulate n_{ysim} values from $f_{Y_i}(\hat{\beta}_Y)$ for $i = 1, \dots, n$, denoted $y_i^{(j)}, j = 1, \dots, n_{ysim}$. Simulation can most efficiently be carried out by drawing one set of n_{ysim} values from the uniform distribution and then using the inverse cumulative distribution method for transformation. Then, calculate

$$\widehat{PS}(\mathbf{Z}_i; \hat{\beta}) = \frac{1}{n_{ysim}} \sum_{j=1}^{n_{ysim}} \hat{P}(X_i=1|\mathbf{Z}_i, Y_i=y_i^{(j)}; \hat{\beta}_X, \hat{\beta}_{X^*}, \hat{\beta}_R). \quad (18)$$

Plugging appropriate estimated quantities into Equations (10) and (11) produces estimates of $PPV(\mathbf{Z}, Y)$ and $NPV(\mathbf{Z}, Y)$, denoted $PP\hat{V}(\mathbf{Z}, Y)$ and $NP\hat{V}(\mathbf{Z}, Y)$, respectively. A step-by-step workflow for multiply imputing M sets of missing X_i and estimating μ_x proceeds as follows:

1. Sample participants with $R_i = 0$ with replacement (bootstrap sample).
2. Obtain $\hat{\beta}_{X^*}$, the maximum likelihood estimate of β_{X^*} , by regressing X^* on $\mathbf{Z}$ and Y (e.g., logistic regression) using the bootstrapped sample.
3. Calculate $\pi_{X_i^*|\mathbf{Z}_i Y_i 0}(\hat{\beta}_{X^*})$ for the original sample with $R = 0$.
4. Plug $\pi_{X_i^*|\mathbf{Z}_i Y_i 0}(\hat{\beta}_{X^*})$ in place of $P(X^* = 1 | \mathbf{Z}, Y, R = 0)$ into Equation (9) to obtain $\hat{P}(X_i = 1 | \mathbf{Z}_i, Y_i, R_i = 0)$ for each participant i with $R_i = 0$. Calculate $PP\hat{V}(\mathbf{Z}_i, Y_i)$ and $NP\hat{V}(\mathbf{Z}_i, Y_i)$ using Equations (10) and (11).
5. For each i with $R_i = 0$, draw a value X_i^{sim} from $[1 - NP\hat{V}(\mathbf{Z}_i, Y_i)] \times (1 - X_i^*) + PP\hat{V}(\mathbf{Z}_i, Y_i) \times X_i^*$.
6. Plug $\hat{P}(\hat{X}_i = 1 | \mathbf{Z}_i, Y_i, R_i = 0)$, $\pi_{R_i|\mathbf{Z}_i Y_i}(\hat{\beta}_R)$, and $\pi_{X_i|\mathbf{Z}_i Y_i 1}(\hat{\beta}_X)$ into Equation (4) to obtain $\hat{P}(\hat{X}_i = 1 | \mathbf{Z}_i, Y_i = y; \hat{\beta}_X, \hat{\beta}_{X^*}, \hat{\beta}_R)$.
7. Solve Equation (17) or (18) to obtain estimated propensity scores, $\widehat{PS}(\mathbf{Z}_i; \hat{\beta})$.
8. Obtain an estimate of $\mu_x, x = 0, 1$ by plugging completed X , $X_i^{comp} = R_i X_i + (1 - R_i) X_i^{sim}$, and $\widehat{PS}(\mathbf{Z}_i; \hat{\beta})$ into a complete-data propensity-score method for standardization such as inverse-probability weighting with weights $W_{X_i|\mathbf{Z}_i} = X_i^{comp} / \widehat{PS}(\mathbf{Z}_i; \hat{\beta}) + (1 - X_i^{comp}) / (1 - \widehat{PS}(\mathbf{Z}_i; \hat{\beta}))$.

Steps 1–2 account for the uncertainty of $\hat{\beta}_{X^*}$, which can alternatively be achieved by simulating from the estimated asymptotic distribution of $\hat{\beta}_{X^*}$ [5, 24]; steps 3–4 compute the imputation model, which is a function of q_{sens} , q_{spec} , and $\pi_{X_i^*|\mathbf{Z}_i Y_i 0}(\hat{\beta}_{X^*})$; step 5 generates the imputations; steps 6–7 compute propensity scores by averaging over R and Y (this can be simplified by regressing X^{comp} on $\mathbf{Z}$, but with the potential for model incompatibility; for

example, a presumed linear-logistic model for $P(X = x | \mathbf{Z})$ may contradict the models specified on the right side of Equation (4)); and step 8 is the complete-data analysis. Repeating steps 1 – 8 M times produces M completed data sets and M estimates of μ_x and corresponding variance-covariance matrix. The final estimate and variance-covariance matrix is obtained using Rubin's combining rules [5].

Although computationally simple, a major drawback of the parametric estimation approach is that it requires specification of $f_{Y_i|\mathbf{Z}_i}(y_i; \beta_Y)$, $\pi_{R_i|\mathbf{Z}_i Y_i}(\beta_R)$, and $\pi_{X_i|\mathbf{Z}_i Y_i 1}(\beta_X)$, which are nuisance parameters that are not of scientific interest, but are needed to obtain estimates of $PS(\mathbf{Z})$ that are compatible with the imputation model. Additionally, even if q_{sens} and q_{spec} are correct, results will not be robust to misspecification of the models in Equations (12)–(15). To overcome this limitation, we consider non-parametric estimation of Equation (13) and of $\widehat{PS}(\mathbf{Z}_i)$, which avoids having to specify and estimate Equations (12), (14), and (15).

3.2. Non-parametric Estimation

We consider machine-learning methods for imputing missing X and for estimating propensity scores by regressing X^{comp} on $\mathbf{Z}$. Rather than positing a data-generating model as in the parametric estimation procedure above, machine-learning methods seek to extract the relationships between an outcome and set of predictor variables without a presumed data-generating model. Machine-learning methods are potentially beneficial for the problem of misclassified proxy data owing to the need for $\pi_{X_i^*|\mathbf{Z}_i Y_i 0}$ to calculate $PPV(\mathbf{Z}, Y)$ and $NPV(\mathbf{Z}, Y)$ for imputation and $PS(\mathbf{Z})$ for analysis. Specifically, these methods allow us to avoid modeling Equations (12), (14) and (15) altogether by eliminating the need to explicitly solve Equations (4) and (17) or (18) to compute $PS(\mathbf{Z})$ that ensure compatibility with the imputation model.

Multiple machine-learning methods have been studied for estimating propensity scores, including CART and ensembles of CARTs such as bagged CART, boosted CART, and random forests [31]. We will only consider boosted CART here for brevity and because it has empirically demonstrated better performance than other machine-learning methods [31]. Briefly, CART recursively partitions the data into nodes defined by a set of predictors and predictor cut points within which observations have similar outcomes. The result is a decision tree that can accommodate interactions and nonlinear relationships. However, CART is prone to over fitting and is suboptimal at revealing linear (or other smooth) main effects.

To overcome these weaknesses, boosted CART passes through the data multiple times to update a suboptimal prediction model. The update is a regression tree of the residuals of the current model [9, 32], and fit is quantified using the log-likelihood. Consider the objective of estimating $\pi_{X_i^*|\mathbf{Z}_i Y_i 0}$. Following the notation of McCaffrey *et al* [32], let $g(\mathbf{Z}_i, Y_i)$ denote the current prediction model for the log-odds that $X_i^* = 1$ given $(\mathbf{Z}_i, Y_i, R_i = 0)$ and let $h(\mathbf{Z}_i, Y_i)$ denote an update to the current model. To estimate $h(\mathbf{Z}_i, Y_i)$, calculate residuals of the estimated current model, $\hat{g}(\mathbf{Z}_i, Y_i)$, $resid_i = X_i^* - \text{expit}[\hat{g}(\mathbf{Z}_i, Y_i)]$, where $\text{expit}[\cdot] = \exp(\cdot)/(1 + \exp(\cdot))$, and use CART to regress $resid_i$ on $(\mathbf{Z}_i, Y_i)$. By modeling the residuals, $h(\mathbf{Z}_i, Y_i)$ is interpreted as the expected score function, thus quantifying the optimal adjustment of $g(\mathbf{Z}_i,$

Y_i) to increase the log-likelihood [32]. By using CART to model the residuals, study participants are partitioned into K regions, $T_1, \dots, T_K$, where the within-region residuals are relatively homogeneous (compared to between-region residuals), and where the estimated within-region mean of residuals is a constant. However, updating $\hat{g}(\mathbf{Z}_i, Y_i)$ using the within-region mean residuals is not guaranteed to increase the likelihood. Therefore, as proposed by Friedman [10], the within-region likelihood is maximized via a constant update, θ_k , for $k = 1, \dots, K$. That is, for $i \in T_k$, maximize

$$L(\theta_k) = \sum_{i \in T_k} X_i^* \times [\hat{g}(\mathbf{Z}_i, Y_i) + \theta_k] - \log\{1 + \exp[\hat{g}(\mathbf{Z}_i, Y_i) + \theta_k]\}. \quad (19)$$

Therefore, using a second-order Taylor-series approximation to reduce computation, the estimated update is

$$\hat{\theta}_k = \hat{h}_k(\mathbf{Z}_i, Y_i) \approx \frac{\sum_{i \in T_k} X_i^* - \text{expit}[\hat{g}(\mathbf{Z}_i, Y_i)]}{\sum_{i \in T_k} \text{expit}[\hat{g}(\mathbf{Z}_i, Y_i)] \{1 - \text{expit}[\hat{g}(\mathbf{Z}_i, Y_i)]\}}, \quad (20)$$

and the estimated updated model is $\hat{g}_{new}(\mathbf{Z}_i, Y_i) = \hat{g}(\mathbf{Z}_i, Y_i) + \hat{h}_k(\mathbf{Z}_i, Y_i)$ for $i \in T_k$. A shrinkage parameter, $\alpha \in (0, 1]$, can be used to reduce the size of the update so that $\hat{g}_{new}(\mathbf{Z}_i, Y_i) = \hat{g}(\mathbf{Z}_i, Y_i) + \alpha \times \hat{h}_k(\mathbf{Z}_i, Y_i)$ for $i \in T_k$. Friedman [33] proposed adding a random-sampling step into the estimation algorithm, where, at each iteration, a random sub-sample is used in CART to estimate the update.

The imputation model is then estimated by computing fitted values, $\hat{\pi}_{X_i^* | \mathbf{Z}_i, Y_i, 0}$, which are used to estimate $PPV(\mathbf{Z}, Y)$ and $NPV(\mathbf{Z}, Y)$. Boosted CART is similarly used to estimate $PS(\mathbf{Z}_i)$ by replacing X_i^* with X_i^{comp} and $(\mathbf{Z}_i, Y_i)$ with $\mathbf{Z}_i$ in Equations (19) and (20), and by including all participants irrespective of R_i . The step-by-step workflow to impute M sets of missing X_i and estimate μ_x is:

1. Sample participants with $R_i = 0$ with replacement (bootstrap sample).
2. Perform a boosted CART analysis of X^* on $\mathbf{Z}$ and Y using the bootstrapped sample.
3. Calculate $\hat{\pi}_{X_i^* | \mathbf{Z}_i, Y_i, 0}$ by applying the boosted CART results to the original sample with $R = 0$.
4. Plug $\hat{\pi}_{X_i^* | \mathbf{Z}_i, Y_i, 0}$ in place of $P(X^* = 1 | \mathbf{Z}, Y, R = 0)$ into Equation (9) to obtain $P(\hat{X}_i = 1 | \mathbf{Z}_i, Y_i, R_i = 0)$ for each participant i with $R_i = 0$. Calculate $PPV(\mathbf{Z}_i, Y_i)$ and $NPV(\mathbf{Z}_i, Y_i)$ using Equations (10) and (11).
5. For each i with $R_i = 0$, draw a value, X_i^{sim} , from $[1 - NPV(\mathbf{Z}_i, Y_i)] \times (1 - X_i^*) + PPV(\mathbf{Z}_i, Y_i) \times X_i^*$.
6. Perform boosted CART of $X_i^{comp} = R_i X_i + (1 - R_i) X_i^{sim}$ on $\mathbf{Z}$ to obtain estimated propensity scores, $\widehat{PS}(\mathbf{Z}_i)$.

7. Obtain an estimate of μ_x , $x = 0, 1$ by plugging X_i^{comp} and $\widehat{PS}(Z_i)$ into a complete-data propensity-score method for standardization such as inverse-probability weighting with weights $W_{X_i|Z_i} = X_i^{comp} / \widehat{PS}(Z_i) + (1 - X_i^{comp}) / (1 - \widehat{PS}(Z_i))$.

Steps 1–2 account for the uncertainty of $\pi_{X_i^*|Z_i Y_i 0}$ [5, 24]; steps 3–4 compute the imputation model, a function of q_{sens} , q_{spec} , and $\pi_{X_i^*|Z_i Y_i 0}$; step 5 generates the imputations; step 6 computes propensity scores by regressing X^{comp} on Z (and avoids model incompatibility without requiring models for Equations (12), (14), and (15) owing to the non-parametric estimation); and step 7 is the complete-data analysis. Repeating steps 1–7 M times produces M completed data sets and M estimates of μ_x and corresponding variance-covariance matrix. Once again, we use Rubin's combining rules to obtain the final estimates [5].

3.3. Adaptations and Extensions

The proposed model and estimation procedures provide a foundation for adaptation and extension. For example, specifying non-differential misclassification can be accomplished by estimating $P(X^* = 1 | Z, R = 0)$ or $P(X^* = 1 | R = 0)$ and then computing sensitivity and specificity by plugging it, rather than $P(X^* = 1 | Y, Z, R = 0)$, into Equations (7) and (8). Under this assumption, a model for Y is not needed for parametric estimation of $PS(Z)$.

The model for differential misclassification can be further generalized. As an example, differential misclassification in the special case of binary Y and a single binary Z is often operationalized by specifying four separate values of sensitivity and specificity, one for each combination of Y and Z . The exponential tilt models can accommodate this case by allowing q_{sens} and q_{spec} to be non-negative functions rather than scalars. Setting $q_{sens}(Z, Y) = \gamma_0 + \gamma_1 Z + \gamma_2 Y + \gamma_3 ZY$ and $q_{spec}(Z, Y) = \xi_0 + \xi_1 Z + \xi_2 Y + \xi_3 ZY$ by user-specified γ and ξ is equivalent to specifying (up to) four values of sensitivity and specificity. The analysis for general Z proceeds by computing $Sens(Z, Y)$ and $Spec(Z, Y)$ and plugging them into Equation (9). Up to now, we considered the special case where $q_{sens}(Z, Y) = \gamma_0$ and $q_{spec}(Z, Y) = \xi_0$. The exponential tilt models are beneficial in that they can be generalized to accommodate continuous Y and multiple, possibly continuous, Z . Indexing exponential tilt models with a flexible function is consistent with their use in the missing-data literature [28, 34].

Also, rather than specifying the parameters of $q_{sens}(Z, Y)$ and $q_{spec}(Z, Y)$ to be fixed quantities, the parameters can be treated as random quantities with distributions. For each iteration of multiple imputation, the parameters of $q_{sens}(Z, Y)$ and $q_{spec}(Z, Y)$ can be simulated from a user-specified distribution. Doing so extends the probabilistic sensitivity analysis of misclassification proposed by Fox *et al* [35] that accounts for uncertainty of assumptions about the misclassification mechanism.

Lastly, the proposed approach was motivated by the scenario where some participants have the gold-standard measure (i.e., $R = 1$ for some). An additional benefit of the approach is that it can handle, as a special case, the scenario where $R = 0$ for all. This special-case scenario is the classical case of exposure misclassification that is most often addressed in the misclassification literature [3]. Non-parametric estimation can be directly applied to this case because propensity scores are estimated by regressing X^{comp} on Z . Parametric

estimation can be adapted—and simplified—by making the following small changes: Equations (12) and (14) need not be estimated, Equation (15) and the propensity scores implicitly condition on $R = 0$, and hence propensity scores are computed by integrating Y out of $P(X = 1 \mid \mathbf{Z}, Y, R = 0)$.

4. Simulation Study

We performed a simulation study to evaluate the finite-sample properties of both the parametric and non-parametric estimation procedures. We explored the cases of binary Y and continuous Y . Boosted CART was implemented in R software version 2.15.0 [36] using the *twang* package [37] with a shrinkage parameter of 0.0005, 20000 iterations, and a 50% sub-sampling fraction. We selected these values based on published recommendations [31, 32]. For both types of outcomes and both types of estimation, we assessed the methods' accuracy by calculating percent relative bias. Empirical standard errors were compared to estimated standard errors, and empirical coverage of the 95% confidence interval was calculated.

For all simulations, we simulated 1000 data sets each of size $n = 500$. We simulated three covariates $\mathbf{Z} = (Z_1, Z_2, Z_3)$ where Z_1 followed a Bernoulli(0.5) distribution, and (Z_2, Z_3) followed a bivariate normal distribution with mean $(0, 0.1Z_1)$, variance $(1, 1)$ and covariance 0.5. We took $M = 50$ imputations for all simulations.

4.1. Binary Outcome

Y was simulated from a Bernoulli distribution with $\text{logit}(f_{Y|\mathbf{Z}}(1)) = 0.6 + 0.1Z_1 + 0.1Z_2 - 0.1Z_3$; and R was simulated from a Bernoulli distribution with $\text{logit}(\pi_{R|\mathbf{Z}Y}) = 0.2 - 0.2Z_1 + 0.1Z_2 - 0.1Z_3 + 0.9Y + 0.1Y Z_2$. When $R = 1$, X was simulated from a Bernoulli distribution with $\text{logit}(\pi_{X|\mathbf{Z}Y 1}) = -0.2 - 0.2Z_1 + 0.2Z_2 - 0.1Z_3 + 1.0Y + 0.1Y Z_2$, and when $R = 0$, X^* was simulated from a Bernoulli distribution with $\text{logit}(\pi_{X^*|\mathbf{Z}Y 0}) = 0.1 + 0.2Z_1 - 0.1Z_2 + 0.1Z_3 - 0.3Y + 0.3Y Z_2$. We set $q_{sens} = 2.00$ and $q_{spec} = 0.75$. The median (interquartile range) of sensitivity and specificity under these models was 0.894 (0.876–0.906) and 0.649 (0.619–0.688), respectively. We estimated standardized proportions $p_x \equiv \mu_x$ and their difference, $p_1 - p_0$, using inverse-probability weighting with multiply imputed X and $W_{X|\mathbf{Z}}$ estimated using both parametric modeling (logistic regression) of Equations (12)–(15) and non-parametric modeling (boosted CART) of Equation (13) and $PS(\mathbf{Z})$. Estimation was performed assuming both the correct values of q_{sens} and q_{spec} and under the incorrect assumption that $q_{sens} = q_{spec} = 99$, an arbitrarily large value to approximate ∞ . We also performed analysis using only data with $R = 1$ (“participant-only estimation”) and with $XR + (1 - R)X^*$ as the exposure (“participant + proxy estimation”) to demonstrate the latter approach's equivalence with the proposed approach assuming $q_{sens} = q_{spec} = \infty$.

Table 1 shows that parametric multiple imputation-based estimation with correct q_{sens} and q_{spec} produced proportions and differences of proportions with negligible bias, whereas participant-only estimation produced biased estimates due to Y -dependent MNAR missingness, and participant + proxy estimation produced biased estimates due to misclassification error. Furthermore, the parametric multiple imputation-based estimates with $q_{sens} = q_{spec} = 99$ were nearly identical to those with participant + proxy estimation,

empirically demonstrating the equivalence between these two models. Not surprisingly, multiple imputation over-estimated the standard errors [38, 39], leading to empirical coverage > 0.95 with correct q_{sens} and q_{spec} . Non-parametric multiple imputation produced proportions that were unbiased and differences in proportions that had small bias when q_{sens} and q_{spec} were correctly specified, similar standard errors to those from parametric estimation, and empirical coverage > 0.95 .

4.2. Continuous Outcome

Y was simulated from a normal distribution with $E[Y | \mathbf{Z}] = 0.3 + 0.1Z_1 + 0.1Z_2 - 0.1Z_3$ and variance 1. R , X , and X^* were simulated under the same distributions as those for binary Y above; we similarly set $q_{sens} = 2.00$ and $q_{spec} = 0.75$. The median (interquartile range) of sensitivity and specificity under these models was 0.892 (0.876–0.919), and 0.654 (0.608–0.688), respectively. We estimated μ_0 , μ_1 , and $\mu_1 - \mu_0$ using inverse-probability weighting with multiply imputed X . $W_{X|Z}$ was parametrically estimated using logistic regression for Equations (12)–(14) and linear regression for Equation (15); and non-parametrically estimated using boosted CART for Equation (13) and $PS(\mathbf{Z})$. Estimation was performed assuming the correct values of q_{sens} and q_{spec} and under the incorrect assumption that $q_{sens} = q_{spec} = 99$. We also performed participant-only estimation and participant + proxy estimation.

Table 2 shows results analogous to those for the binary outcome. When q_{sens} and q_{spec} were correctly specified, parametric and non-parametric multiple imputation produced unbiased means and mean differences, but with empirical coverage > 0.95 .

5. Data Application: The Baltimore Hip Studies

We illustrate our proposed statistical methods using data from the Second Cohort of the Baltimore Hip Studies, a prospective study comprising older adults who experienced a hip fracture [11]. The goal of the present analysis was to determine whether perceived rapid recovery of independent mobility, assessed using self-reported ability to walk 10 feet without human assistance two months post hip fracture, is associated with survival two years after the fracture. We considered both a binary outcome and a continuous outcome. For the binary outcome, we operationalized two-year survival as alive or dead; for the continuous outcome, we operationalized survival as the number of days alive in the two years after hip fracture (maximum, 731 days).

The analysis included 502 participants, where 365 provided self-reports, among whom 284 (77.8%) reported independent mobility. Proxies assessed mobility for the remaining 137 participants, 61 (44.5%) of whom evaluated the participant to be independently mobile. All estimated proportions and means were standardized for sex (395 women, 107 men), years of age (mean=80.8, SD=7.3, range 65–104) and number of comorbid conditions (mean=3.2, SD=2.1, range 0–12). Participant-only, participant + proxy, and parametric and non-parametric multiple imputation analyses were all performed using inverse-probability weighting. Multiple imputation was carried out with $M = 50$. Propensity scores from participant-only and participant + proxy analyses were estimated using conventional logistic regression to realistically reflect a typical analysis that does not account for missingness or

misclassification. Boosted CART was implemented using a shrinkage parameter of 0.0005, 20000 iterations, and a 50% sub-sampling fraction, as per published recommendations [31, 32]. Balance diagnostics demonstrated that these tuning parameters performed well.

5.1. Binary Outcome

There were 404 (80.5%) participants alive two years after hip fracture. We performed multiple imputation assuming $(q_{sens}, q_{spec}) = (3, 1)$ and $(q_{sens}, q_{spec}) = (1, 3)$. Using a parametric model, assuming $(q_{sens}, q_{spec}) = (3, 1)$ produced sensitivity and specificity distributions with median (interquartile range) of 0.950 (0.929–0.960) and 0.744 (0.692–0.806), respectively; assuming $(q_{sens}, q_{spec}) = (1, 3)$ produced sensitivity and specificity distributions with median (interquartile range) of 0.718 (0.640–0.767) and 0.955 (0.943–0.968), respectively. Using boosted CART, setting $(q_{sens}, q_{spec}) = (3, 1)$ produced sensitivity and specificity distributions with median (interquartile range) of 0.937 (0.903–0.965) and 0.785 (0.665–0.855), respectively; setting $(q_{sens}, q_{spec}) = (1, 3)$ produced sensitivity and specificity distributions with median (interquartile range) of 0.669 (0.556–0.783) and 0.964 (0.938–0.978), respectively. We also performed analysis assuming $(q_{sens}, q_{spec}) = (99, 99)$.

Table 3 shows that standardized two-year survival was higher for independently mobile participants than for participants who were not independently mobile, but the magnitude of the difference varied by statistical method. Among the methods and assumptions considered, boosted CART with $(q_{sens}, q_{spec}) = (99, 99)$ and with $(q_{sens}, q_{spec}) = (1, 3)$ produced the largest and smallest differences in survival, 0.139 and 0.110, respectively. The participant-only analysis produced the highest estimated proportions of survival in both groups, and estimates from the participant + proxy analysis and parametric multiple imputation with $q_{sens} = q_{spec} = 99$ were nearly identical.

We additionally performed multiple imputation using 100 combinations of q_{sens} and q_{spec} ranging from 0.01 to 4 and 99. When these broad ranges of q_{sens} and q_{spec} were considered, their effects on estimated proportions could be discerned using both parametric and non-parametric estimation. p_0 ranged 0.713–0.754 and 0.707–0.741 using parametric and non-parametric estimation, respectively. Similarly, p_1 ranged 0.812–0.874 and 0.811–0.870 for parametric and non-parametric estimation, respectively. As a result, $p_1 - p_0$ ranged 0.058–0.154 and 0.070–0.145 for parametric and non-parametric estimation, respectively.

5.2. Continuous Outcome

Participants lived for an average of 659 (SD=174) days after hip fracture. We considered the same sets of values for (q_{sens}, q_{spec}) as we did for binary survival. Using continuous survival with a parametric model, setting $(q_{sens}, q_{spec}) = (3, 1)$ produced sensitivity and specificity distributions with median (interquartile range) of 0.947 (0.922–0.957) and 0.754 (0.708–0.822), respectively; setting $(q_{sens}, q_{spec}) = (1, 3)$ produced sensitivity and specificity distributions with median (interquartile range) of 0.706 (0.614–0.753) and 0.958 (0.947–0.972), respectively. To address the large proportion alive after two years, we calculated propensity scores assuming that Y follows the density $I(Y = 731)P(Y = 731 | \mathbf{Z}) + I(Y < 731)P(Y < 731 | \mathbf{Z})f(Y = y | Y < 731, \mathbf{Z})$, where $f(\cdot)$ is assumed to be a normal density. Using boosted CART, setting $(q_{sens}, q_{spec}) = (3, 1)$ produced sensitivity and specificity

distributions with median (interquartile range) of 0.943 (0.904–0.966) and 0.767 (0.660–0.852), respectively; setting $(q_{sens}, q_{spec}) = (1, 3)$ produced sensitivity and specificity distributions with median (interquartile range) of 0.690 (0.564–0.788) and 0.961 (0.936–0.977), respectively.

Results in Table 4 show that standardized two-year life expectancy was higher for independently mobile participants than for participants who were not independently mobile, but the magnitude of the difference varied by statistical method. Among the methods and assumptions considered, parametric analysis with $(q_{sens}, q_{spec}) = (99, 99)$ and participant-only analysis produced the largest and smallest differences in average two-year life expectancy, 74 and 44 days, respectively. The participant-only analysis produced the highest average two-year life expectancy in both groups. Estimates from the participant + proxy analysis and parametric multiple imputation with $q_{sens} = q_{spec} = 99$ differed, owing to differences in how the propensity score was calculated; however, estimates from participant + proxy analysis were similar to those from boosted CART with $q_{sens} = q_{spec} = 99$.

When assessing broader ranges of q_{sens} and q_{spec} (0.01 to 4 and 99), we found that $\hat{\mu}_0$ ranged 606–645 days and 616–639 days for parametric and non-parametric estimation, respectively. Similarly, $\hat{\mu}_1$ ranged 660–686 days and 660–684 days for parametric and non-parametric estimation, respectively. In this case, $\hat{\mu}_1 - \hat{\mu}_0$ ranged 14–76 days and 20–64 days for parametric and non-parametric estimation, respectively.

6. Discussion

This paper proposed and evaluated statistical methods to address exposures that are missing and assessed using error-prone proxy reports for analysis with both categorical and continuous outcomes. A major innovation of the parametric modeling approach is the use of pattern-mixture models to address missing and differentially misclassified exposure data. Pattern-mixture models have hitherto only been used to address missing or differentially misclassified outcomes [7, 22–24]. This approach was made possible by 1) using a likelihood that did not require a model for Y conditioned on X and X^* , and 2) estimating covariate-standardized outcome means. The advantages of this approach are that the models for imputation and analysis are compatible (called ‘congenial’ in the multiple imputation literature [40]) and that the methods can be easily implemented. The disadvantage of the parametric approach is that it requires correctly specifying and estimating models for $f_{Y_i|Z_i}(y; \beta_Y)$ (including the distribution), $\pi_{R_i|Z_iY_i}(\beta_R)$, and $\pi_{X_i|Z_iY_i}(\beta_X)$, which are nuisance parameters. The non-parametric machine-learning approach overcame the limitations of the parametric approach by not requiring specification of the distribution for Y and circumventing estimation of nuisance parameters while preserving model compatibility.

Our proposed approach incorporates features from and generalizations of other methods that were developed to handle exposure misclassification in the case where $R = 0$ for all participants. Lyles and Lin [26] proposed predictive-value weighting for logistic regression with binary outcomes and jackknife estimation of standard errors. The proposed approach uses the predictive values for multiple imputation, a method that can decrease the computational burden of estimating standard errors (i.e., 50 imputations rather than n

jackknife iterations). Fox *et al* [35] proposed reconstructing the data that would have been observed had there been no misclassification using Monte Carlo simulations from predictive values and then performing logistic regression on the completed data; however, the authors' predictive values did not accommodate covariates, and outcomes were assumed to be binary. Our proposed parametric estimation uses a likelihood decomposition that ensures all specified models are compatible with each other; namely, the propensity score model is compatible with the predictive-value models. Our proposed non-parametric approach ensures model compatibility by requiring fewer models—the same number of models as the proposed approaches by Lyles and Lin [26] and by Fox *et al* [35]. A direct adaptation of these previous methods in which a linear-logistic model is presumed for $P(X^* = 1 \mid \mathbf{Z}, Y, R = 0)$ and a linear-logistic (binary outcome) or linear (continuous outcome) model is presumed for $E(Y \mid X = x, \mathbf{Z} = \mathbf{z})$ could result in model incompatibility, because the presumed links and linear relationships may not be preserved. That is, a full-data distribution that is consistent with all model assumptions may not exist.

In summary, our approach generalizes these earlier proposals by 1) considering the case where some participants have the gold-standard rather than error-prone exposure measured (i.e., $R = 1$ for some), 2) ensuring compatibility of all specified models and compatibility of assumptions about sensitivity and specificity with the data, 3) accommodating continuous or binary outcomes, 4) using standardization to facilitate the analyst's choice of association measure (e.g., $\mathbf{Z}$ -adjusted risk difference, risk ratio, or odds ratio for binary outcomes), and 5) non-parametric estimation with machine-learning methods to further prevent model misspecification.

The proposed method is similar in spirit to our published method for handling error-prone proxy-reported outcomes [7], but the new method features multiple innovations to help overcome some challenges of misclassified exposures that are not encountered in analysis with misclassified outcomes. In particular, unlike with misclassified outcomes, having misclassified exposures may require modeling both the outcome and the exposure. The propensity-score approach with non-parametric estimation is of benefit because it circumvents the need for an outcome model. Furthermore, when misclassification is differential, the exposure misclassification model (the imputation model) must condition on the outcome, and this model may not be compatible with the analysis model (an outcome model or propensity-score model). Therefore, we proposed a novel likelihood decomposition for parametric estimation and evaluated a non-parametric estimation procedure. A propensity-score method that can handle differential exposure misclassification is an innovation in itself. While methods for covariate measurement error in propensity-score analysis are available [41–43], this is the first method, to our knowledge, that handles exposure misclassification in propensity-score analysis. The proposed method can be adapted to handle the situation where all participants have the error-prone “proxy” exposure measure as a special case ($R = 0$ for all). A recent simulation study demonstrating the biasing effects of exposure misclassification on propensity-score estimators supports the need for such methods [44].

Both parametric and non-parametric estimation require user-specified values that encode presumed sensitivity and specificity of proxy reports. However, we do not consider this to

be a limitation. Rather, this feature accurately represents the realities of exposure misclassification, namely that the sensitivity and specificity of proxy reports are not identifiable from the data, and an assumption is needed for estimation. Thus, a strength of the overall approach is that the assumptions are made explicit. In particular, when using the same model to estimate propensity scores, participant + proxy analysis is equivalent to the proposed method with large q_{sens} and q_{spec} .

The proposed methods were motivated by aging research, where proxy data are routinely collected and have been evaluated in published proxy-participant validation studies (see [1, 25] and references therein). However, the validation studies are imperfect because they only generalize to participants who do not need a proxy respondent. Despite this limitation, researchers may have a better intuition about misclassification than about missingness, thus making proxy data a valuable part of a sensitivity analysis. In future work, we aim to extend our methods to formally handle internal validation data as part of a sensitivity analysis for missing and misclassified exposures.

Acknowledgments

Contract/grant sponsor: National Institutes of Health K25AG034216, R01AG041202

References

1. Gruber-Baldini, AL.; Shardell, M.; Lloyd, K.; Magaziner, J. Use of proxies and informants. In: Newman, AB.; Cauley, JA., editors. *The Epidemiology of Aging*. New York: Springer; 2012. p. 81-90.
2. Nelson LM, Longstreth WT Jr, Koepsell TD, Van Belle G. Proxy respondents in epidemiologic research. *Epidemiologic Reviews*. 1990; 12:71–86. [PubMed: 2286227]
3. Carroll, RJ.; Ruppert, D.; Stefanski, LA.; Crainiceanu, CM. *Measurement Error in Nonlinear Models: A Modern Perspective*. 2. London: Chapman and Hall; 2006.
4. Guolo A. Robust techniques for measurement error correction: a review. *Statistical Methods in Medical Research*. 2008; 17:555–580.10.1177/0962280207081318 [PubMed: 18375458]
5. Rubin, DB. *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley; 2004.
6. Shardell M, Hicks GE, Miller RR, Langenberg P, Magaziner J. Pattern-mixture models for analyzing normal outcome data with proxy respondents. *Statistics in Medicine*. 2010; 29:1522–1538.10.1002/sim.3902 [PubMed: 20535763]
7. Shardell M, Simonsick E, Hicks GE, Resnick B, Ferrucci L, Magaziner J. Sensitivity analysis for nonignorable missingness and outcome misclassification from proxy reports. *Epidemiology*. 2013; 24:215–223.10.1097/EDE.0b013e31827f4fa9 [PubMed: 23348065]
8. Breiman, L.; Friedman, JH.; Olshen, RA.; Stone, CJ. *Classification and Regression Trees*. Belmont, CA: Wadsworth International; 1984.
9. Friedman JH, Hastie T, Tibshirani R. Additive logistic regression: a statistical view of boosting (with discussion). *Annals of Statistics*. 2000; 28:337–407.
10. Friedman JH. Greedy function approximation: a gradient boosting machine. *Annals of Statistics*. 2001; 29:1189–1232.
11. Magaziner J, Hawkes W, Hebel JR, Zimmerman SI, Fox KM, Dolan M, Felsenthal G, Kenzora J. Recovery from hip fracture in eight areas of function. *Journals of Gerontology Series A: Biological Sciences Medical Sciences*. 2000; 55:M498–M507.10.1093/gerona/55.9
12. Snow AL, Cook KF, Lin PS, Morgan RO, Magaziner J. Proxies and other external raters: methodological considerations. *Health Services Research*. 2005; 40:1676–1693.10.1111/j.1475-6773.2005.00447.x [PubMed: 16179002]

13. Robins JM. A new approach to causal inference in mortality studies with sustained exposure periods-application to control of the healthy worker survivor effect. *Mathematical Modelling*. 1986; 7:1393–1512.
14. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika*. 1983; 70:41–55.10.1093/biomet/70.1.41
15. Rosenbaum PR, Rubin DB. Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association*. 1984; 79:516–524.10.1080/01621459.1984.10478078
16. D'Agostino RB. Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Statistics in Medicine*. 1998; 17:2265–2281.10.1002/(SICI)1097-0258(19981015)17:19<2265::AID-SIM918>3.0.CO;2-B [PubMed: 9802183]
17. Robins JM, Hernán MA, Brumback B. Marginal structural models and causal inference in epidemiology. *Epidemiology*. 2000; 11:550–560.10.1097/00001648-200009000-00011 [PubMed: 10955408]
18. Sato T, Matsuyama Y. Marginal structural models as a tool for standardization. *Epidemiology*. 2003; 14:680–686. EDE.0000081989.82616.7d. [PubMed: 14569183]
19. Rubin DB. Inference and missing data. *Biometrika*. 1976; 63:581–592.10.1093/biomet/63.3.581
20. Daniel RM, Kenward MG, Cousens SN, De Stavola BL. Using causal diagrams to guide analysis in missing data problems. *Statistical Methods in Medical Research*. 2012; 21:243–256.10.1177/0962280210394469 [PubMed: 21389091]
21. Westreich D. Berkson's bias, selection bias, and missing data. *Epidemiology*. 2012; 23:159–164.10.1097/EDE.0b013e31823b6296 [PubMed: 22081062]
22. Little RJ. Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Association*. 1993; 88:125–134.10.1080/01621459.1993.10594302
23. Little RJ, Wang Y. Pattern-mixture models for multivariate incomplete data with covariates. *Biometrics*. 1996; 52:98–111. [PubMed: 8934587]
24. Little, RJA.; Rubin, DB. *Statistical Analysis with Missing Data*. 2. New York: Wiley; 2002.
25. 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. *Journal of Aging and Health*. 2012; 24:367–383.10.1177/0898264311424208 [PubMed: 22210805]
26. Lyles RH, Lin J. Sensitivity analysis for misclassification in logistic regression via likelihood methods and predictive value weighting. *Statistics in Medicine*. 2010; 29:2297–2309.10.1002/sim.3971 [PubMed: 20552681]
27. Barndorff-Nielsen, OE.; Cox, DR. *Asymptotic Techniques for use in Statistics*. London: Chapman and Hall; 1989.
28. Shardell M, El-Kamary SS. Sensitivity analysis of informatively coarsened data using pattern mixture models. *Journal of Biopharmaceutical Statistics*. 2009; 19:1018–1038.10.1080/10543400903242779 [PubMed: 20183462]
29. 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. *American Journal of Epidemiology*. 2008; 168:1460–1469.10.1093/aje/kwn265 [PubMed: 18952850]
30. Barron BA. The effects of misclassification on the estimation of relative risk. *Biometrics*. 1977; 33:414–418. [PubMed: 884199]
31. Lee BK, Lessler J, Stuart EA. Improving propensity score weighting using machine learning. *Statistics in Medicine*. 2010; 29:337–346.10.1002/sim.3782 [PubMed: 19960510]
32. McCaffrey DF, Ridgeway G, Morral AR. Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological Methods*. 2004; 4:403–425.10.1037/1082-989X.9.4.403 [PubMed: 15598095]
33. Friedman JH. Stochastic gradient boosting. *Computational Statistics and Data Analysis*. 2002; 38:367–378.10.1016/S0167-9473(01)00065-2

34. 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; 65:275–297.10.1111/1467-9868.00386
35. Fox MP, Lash TL, Greenland S. A method to automate probabilistic sensitivity analyses of misclassified binary variables. *Int J Epidemiol*. 2005; 34:1370–1376.10.1093/ije/dyl226 [PubMed: 16172102]
36. Ihaka R, Gentleman R. R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics*. 1996; 5:299–314.
37. Ridgeway G, McCaffrey DF, Morral AR. Twang: Toolkit for weighting and analysis of nonequivalent groups. R Package Version 1.0-1. 2006
38. Wang N, Robins JM. Large-sample theory for parametric multiple imputation procedures. *Biometrika*. 1998; 85:935–948.10.1093/biomet/85.4.935
39. Robins JM, Wang N. Inference for imputation estimators. *Biometrika*. 2000; 87:113–124.10.1093/biomet/87.1.113
40. Meng XL. Multiple-imputation inferences with uncongenial sources of input. *Statistical Science*. 1994; 9:538–558.10.1214/ss/1177010269
41. McCaffrey DF, Lockwood JR, Setodji CM. Inverse probability weighting with error-prone covariates. *Biometrika*. 2013; 100:671–680.10.1093/biomet/ast022 [PubMed: 24795484]
42. Yi GY, Ma Y, Carroll RJ. A functional generalized method of moments approach for longitudinal studies with missing responses and covariate measurement error. *Biometrika*. 2012; 99:151–165.10.1093/biomet/asr076
43. D’Agostino R Jr, Rubin DB. Estimating and using propensity scores with partially missing data. *J Am Statist Assoc*. 2000; 95:749–759.10.1080/01621459.2000.10474263
44. Babanezhad M, Vansteelandt S, Goetghebeur E. Comparison of causal effect estimators under exposure misclassification. *Journal of Statistical Planning and Inference*. 2010; 140:1306–1319.10.1016/j.jspi.2009.11.015

Table 1

Simulation Study Results for Binary Outcome.^a

Parameter	Analytic Method	Estimate	% Relative Bias	SE	ESE	Coverage ^b
p_1	Participant Only ^c	0.557	22	0.040	0.040	27.4
	Participant + Proxy ^d	0.419	-8	0.030	0.031	76.5
	Parametric MI, Correct ^e	0.456	<1	0.033	0.032	96.1
	Parametric MI, Incorrect ^f	0.418	-8	0.030	0.031	76.5
p_0	Non-parametric MI, Correct ^g	0.457	<1	0.034	0.030	97.5
	Non-parametric-MI, Incorrect ^h	0.419	-8	0.031	0.031	77.7
	Participant Only	0.316	10	0.040	0.041	89.2
	Participant + Proxy	0.300	5	0.030	0.030	91.3
	Parametric MI, Correct	0.287	<1	0.027	0.026	95.5
	Parametric MI, Incorrect	0.300	5	0.030	0.032	91.6
$p_1 - p_0$	Non-parametric MI, Correct	0.285	<1	0.028	0.026	95.5
	Non-parametric-MI, Incorrect	0.300	5	0.031	0.032	92.7
	Participant Only	0.242	43	0.056	0.058	73.5
	Participant + Proxy	0.119	-30	0.043	0.044	77.2
	Parametric MI, Correct	0.169	<1	0.043	0.039	97.2
	Parametric MI, Incorrect	0.119	-30	0.043	0.045	77.1
	Non-parametric MI, Correct	0.172	2	0.044	0.037	98.3
	Non-parametric-MI, Incorrect	0.120	-29	0.044	0.044	79.5

^a1000 simulations, 500 observations. True parameters: $p_0 = 0.287$, $p_1 = 0.455$, $p_1 - p_0 = 0.169$.^bPercent of 95% confidence intervals covering true parameter values;^cParticipant Only = Analysis excluding observations with $R = 0$;^dParticipant + Proxy = Analysis substituting missing X with X*;^eParametric MI, Correct = Multiple imputation with imputation model and propensity scores estimated using maximum likelihood correctly assuming $q_{sens} = 2.00$ and $q_{spec} = 0.75$;^fParametric MI, Incorrect = Parametric MI incorrectly assuming $q_{sens} = q_{spec} = 99$;

g Non-parametric MI, Correct = Multiple imputation with imputation model and propensity scores estimated using boosted CART correctly assuming $q_{sens} = 2.00$ and $q_{spec} = 0.75$;

h Non-parametric MI, Incorrect = Non-Parametric MI incorrectly assuming $q_{sens} = 99$ and $q_{spec} = 99$.

Table 2

Simulation Study Results for Continuous Outcome.^a

Parameter	Analytic Method	Estimate	% Relative Bias	SE	ESE	Coverage ^b
μ_1	Participant Only ^c	0.702	41	0.052	0.050	2.4
	Participant + Proxy ^d	0.423	-15	0.045	0.045	61.4
	Parametric MI, Correct ^e	0.499	<1	0.048	0.043	96.6
	Parametric MI, Incorrect ^f	0.423	-15	0.045	0.044	61.2
μ_0	Non-Parametric MI, Correct ^g	0.501	<1	0.049	0.044	97.2
	Non-Parametric-MI, Incorrect ^h	0.430	-14	0.046	0.046	67.7
	Participant Only	0.285	40	0.057	0.057	69.6
	Participant + Proxy	0.238	17	0.044	0.044	88.9
	Parametric MI, Correct	0.202	<1	0.041	0.038	96.9
	Parametric MI, Incorrect	0.238	17	0.044	0.044	88.9
$\mu_1 - \mu_0$	Non-Parametric MI, Correct	0.202	<1	0.042	0.037	97.1
	Non-Parametric-MI, Incorrect	0.238	17	0.045	0.044	88.3
	Participant Only	0.417	41	0.077	0.074	65.6
	Participant + Proxy	0.186	-37	0.063	0.063	58.0
	Parametric MI, Correct	0.297	<1	0.063	0.051	98.6
	Parametric MI, Incorrect	0.186	-37	0.063	0.063	58.4
	Non-Parametric MI, Correct	0.299	<1	0.064	0.054	98.4
	Non-Parametric-MI, Incorrect	0.192	-35	0.064	0.064	61.6

^a1000 simulations, 500 observations. True parameters: $\mu_0 = 0.203$, $\mu_1 = 0.500$, $\mu_1 - \mu_0 = 0.296$.^bPercent of 95% confidence intervals covering true parameter values;^cParticipant Only = Analysis excluding observations with $R = 0$;^dParticipant + Proxy = Analysis substituting missing X with X*;^eParametric MI, Correct = Multiple imputation with imputation model and propensity scores estimated using maximum likelihood correctly assuming $q_{sens} = 2.00$ and $q_{spec} = 0.75$;^fParametric MI, Incorrect = Parametric MI incorrectly assuming $q_{sens} = q_{spec} = 99$;

g Non-Parametric MI, Correct = Multiple imputation with imputation model and propensity scores estimated using boosted CART correctly assuming $q_{sens} = 2.00$ and $q_{spec} = 0.75$;

h Non-Parametric MI, Incorrect = Non-Parametric MI incorrectly assuming $q_{sens} = 99$ and $q_{spec} = 99$.

Table 3
Results for 502 Patients in the Second Cohort of the Baltimore Hip Studies, Binary Outcome.

Proportion Alive 2 Years Post-Fracture										
Analytic Method	q_{sens}	q_{spec}	Independently Mobile (IM) ^a			Not Independently Mobile (NIM) ^a			Difference	
			$\hat{p}_{IM}$	SE		$\hat{p}_{NIM}$	SE			
Participant Only ^b			0.88	0.02		0.75	0.05		0.12	0.06
Participant + Proxy ^c			0.85	0.02		0.72	0.04		0.13	0.04
Parametric MI ^d	3	1	0.85	0.02		0.72	0.03		0.13	0.04
	1	3	0.83	0.02		0.72	0.04		0.11	0.04
	99	99	0.85	0.02		0.72	0.04		0.13	0.04
Non-Parametric MI ^e	3	1	0.85	0.02		0.72	0.03		0.13	0.04
	1	3	0.83	0.02		0.72	0.04		0.11	0.05
	99	99	0.85	0.02		0.71	0.04		0.14	0.04

^aMobility assessed two months after hip fracture.
^bParticipant Only = Analysis excluding observations with missing participant self reports;
^cParticipant + Proxy = Analysis substituting missing participant self reports with proxy reports;
^dParametric MI = Multiple imputation using maximum likelihood to estimate the imputation and propensity-score models;
^eNon-Parametric MI =Multiple imputation using boosted CART to estimate the imputation and propensity-score models.

Table 4
Results for 502 Patients in the Second Cohort of the Baltimore Hip Studies, Continuous Outcome.

Mean Days Alive through 2 Years (max, 731) Post-Fracture										
Analytic Method	q_{sens}	q_{spec}	Independently Mobile (IM) ^a			Not Independently Mobile (NIM) ^a			Difference $\hat{\mu}_{IM} - \hat{\mu}_{NIM}$	
			$\hat{\mu}_{IM}$	SE		$\hat{\mu}_{NIM}$	SE			
Participant Only ^b			688	8		644	22		44	24
Participant + Proxy ^c			680	8		618	17		62	19
Parametric MI ^d	3	1	683	8		612	17		71	19
	1	3	673	8		617	19		56	20
	99	99	682	8		608	18		74	20
Non-Parametric MI ^e	3	1	679	9		624	16		55	18
	1	3	671	8		625	18		46	20
	99	99	680	8		619	17		61	19

^aMobility assessed two months after hip fracture.

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

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

^dParametric MI = Multiple imputation using maximum likelihood to estimate the imputation and propensity-score models;

^eNon-Parametric MI = Multiple imputation using boosted CART to estimate the imputation and propensity-score models.