Published in final edited form as:
Biostatistics & Epidemiology (2020), **4**(1): 247–264
DOI: 10.1080/24709360.2020.1794705

Regression with incomplete multivariate surrogate responses for a latent covariate

HUA SHEN

Department of Mathematics and Statistics,
University of Calgary, Calgary, AB, T2N 1N4, Canada
E-mail: hua.shen@ucalgary.ca

RICHARD J. COOK

Department of Statistics and Actuarial Science,
University of Waterloo, Waterloo, ON, N2L 3G1, Canada

Summary

We consider the setting in which a categorical exposure variable of interest can only be measured subject to misclassification via surrogate variables. These surrogate variables may represent the classification of an individual via imperfect diagnostic tests. In such settings, a random number of diagnostic tests may be ordered at the discretion of a treating physician with the decision to order further tests made in a sequential fashion based on the results of preliminary test results. Because the underlying latent status is not ascertainable these cheaper but imperfect surrogate test results are used in lieu of the definitive classification in a model for a long-term outcome. Naive use of a single surrogate or functions of the available surrogates can lead to biased estimators of the association and invalid inference.Wepropose a likelihood-based approach for modeling the effect of the latent variable in the absence of validation data with estimation based on an expectation–maximization (EM) algorithm. The method yields consistent and efficient estimates and is shown to out-perform several common alternative approaches. The performance of the proposed method is demonstrated in simulation studies and its utility is illustrated by applying the proposed method to the stimulating study on breast cancer.

Keywords: expectation-maximization algorithm, latent variable, misspecified model, regression, surrogate variable

1 INTRODUCTION

Many scientific questions in medical research involve examining the association between an individual's disease status, possibly given other covariates, on a longer-term response. Often, however, the precise disease status can be difficult to determine and hence may be subject to misclassification. Examples arise in psychiatry where mental disorders are difficult to classify (Grove et al., 1981), rheumatology where joint and bone destruction can arise from different kinds of arthritis (Perrot et al., 2012), and radiography where there can be poor images which made definitive diagnoses challenging (Meade et al., 2001). Approaches for dealing with misclassified covariates are well established (Gustafson, 2003; Carroll et al., 2006; Buonaccorsi, 2010; Yi, 2016) but most methods require the use of either a validation study (Greenland, 1988; Marshall, 1990; Holcroft and Spiegelman, 1999; Morrissey and Spiegelman, 1999; Spiegelman et al., 2000) or a replication study (Rindskopf and Rindskopf, 1986; Liu and Liang, 1991; Chu et al., 2009; Yi and He, 2017).

We consider the setting in which there is no formal validation or replication study in the traditional sense. In the setting of interest several diagnostic tests, each with non-zero false positive and false negative error rates, may be ordered in a sequential stochastic manner. The outcomes of prior diagnostic test results influence the probability that further diagnostic tests are ordered. Since they are error-prone, naive use of any one of the test results in lieu of the true disease status, or use of a classification rule based on the collection of error-prone test results, can lead to biased estimates of the association between disease and the response. To address this, we jointly model the latent variable for the disease status, the misclassification rates of the imperfect diagnostic tests, and the response of interest. This multi-part model can lead to insight into the prevalence of the disease, the operating characteristics (i.e. sensitivity and specificity) of the different diagnostic tests, and consistent estimates of the effects of interest.

The proposed method involves the conceptualization of a complete data set which contains information on the latent covariate for the disease status; while this may be categorical we focus on the case of a latent binary covariate. Conditional independence assumptions are carefully described along with assumptions about the mechanism leading to the availability of the surrogate variables based on the imperfect diagnostic test results. An expectation-maximization (EM) algorithm (Dempster et al., 1977) is developed to facilitate estimation, where the maximization step can be easily implemented using standard software provided that it can accommodate weights. This method yields consistent and efficient estimates of the covariate effect, the operating characteristics of the diagnostic tests, and the prevalence of the disease.

We organize the remainder of the paper as follows. In Section 2, we define the notation, write the complete data likelihood function, and describe the details of the expectation step (E-step) and maximization step (M-step) of the EM algorithm. The formulation can handle the setting in which some diagnostic test results are not observed in some individuals. We also describe an alternative two-stage estimation approach in which the first stage is directed at estimation of what may be viewed as nuisance parameters related to the diagnostic test performance and the probability model for the latent disease status. In Section 3, we assess the empirical performances of the estimators arising from possible naive analysis along with those from the proposed method. In Section 4, we provide an illustrative application on breast cancer data. Concluding remarks and topics for further research are provided in Section 5.

2 Likelihood, Estimation and Inference

2.1 General Model Formulation

In what follows, we omit the subscript i indexing individuals and consider the contributions from a single individual. Let Y denote a response of interest and consider the aim to model $E(Y|Z, W)$ where the response Y is an outcome observed after some specified period of follow-up, Z is a discrete latent variable of prime interest which is defined at a baseline assessment, and W are additional fixed covariates which are always observed and measured without error. The variable Z is taken to be categorical here and may represent, for example, an indicator of the presence of a disease, or the extent of damage resulting from disease. Such variables are often latent when it is difficult to diagnose individuals or when the definitive diagnostic test is prohibitively expensive for routine use. We consider the case in which there are potentially K surrogate variables available; we refer to these as surrogates for Z since they may be helpful in inferring the latent state Z. They may represent the results of screening tests, different physicians' assessments, or random variables reflecting the results of any other error-prone attempt at diagnosis. Let X_k denote the kth such variable, $k = 1, \ldots, K$ and $X = (X_1, \ldots, X_K)'$ denote the full $K \times 1$ vector of potential surrogates for Z. Here, we assume X and W do not share any elements; that is, while the latent Z and W may be associated and this may induce an association between X and W, the model for X is specified conditionally only on the latent value of Z under the assumption $X \perp W|Z$.

We consider the setting where the decision to take the particular surrogate measurements (i.e. the elements of X) is made based on available information at the point when the covariates are available. Since some elements of X may be unobserved we let $R_k = I(X_k \text{ is observed})$, define $R = (R_1, \ldots, R_K)'$, and we use X° to denote the observed elements of X and X^m for the unobserved elements of X. The complete data is denoted by $D = \{R, Y, Z, X, W\}$ where Z and all elements of X are observed, and $D^\circ = \{R, Y, X^\circ, W\}$ denotes the observed data.

Since W is always observed we can condition on it and define the observed data likelihood for a generic individual as $L \propto P(R, Y, X^\circ|W)$ which can be written using a selection model factorization as

$$L = \sum_z P(R|Y, Z, X^\circ, W)P(Y|Z, X^\circ, W)P(Z, X^\circ|W) \ . \tag{1}$$

We next lay out some conditional independence assumptions which lead to the models and likelihoods of interest.

Assumption A1: $R \perp Z, X^m|Y, X^\circ, W$.

Assumption A1 allows us to factor the term $P(R|Y, Z, X^\circ, W)$ out of the sum in (1). This is a reasonable assumption in the present setting since the decision to record values for additional elements of X will typically be based on observable quantities recorded in the available X_k terms and W. We discuss this further in Section 2.2 where we consider a further simplification. We also assume the missing data process is non-informative in the sense that it shares no parameters with the models of interest. Taken together these assumptions enable one to avoid modeling the missing data process.

Assumption A2 $Y \perp X|Z, W$.

Assumption A2 implies that if Z and W were observed, then X would convey no additional useful information for the model $Y|Z, W$; this reflects the fact that X may be viewed as surrogates for Z.

Under these assumptions, we can focus on the partial likelihood obtained from (1) and given by

$$L \propto \sum_z P(Y|Z, W)P(X^\circ|Z, W) \cdot P(Z|W) \ , \tag{2}$$

which can be expressed alternatively as

$$L \propto E_Z\{P(Y|Z,W)|X^\circ,W\}P(X^\circ|W)\ . \tag{3}$$

The complete data log-likelihood $\ell = \log \mathcal{L}$ corresponding to (2) is

$$\sum_z I(Z=z)\left[\log P(Y|Z=z,W) + \log P(X^\circ|Z=z,W) + \log P(Z=z|W)\right]\ . \tag{4}$$

We let β index the response model $P(Y|Z,W)$, ξ index the surrogate variable model $P(X|Z,W)$, ζ index the model for the latent state $P(Z|W)$, and $\theta = (\beta',\xi',\zeta')'$. The contribution to the observed data score vector from an individual can be written as

$$S(\theta;\theta) = (S_1'(\beta;\theta), S_2'(\xi;\theta), S_3'(\zeta;\theta))'\ ,$$

where

$$\begin{aligned} S_1(\beta;\theta) &= E_Z\left\{\partial \log P(Y|Z,W;\beta)/\partial\beta|D^\circ;\theta\right\}\ , \\ S_2(\xi;\theta) &= E_Z\left\{\partial \log P(X^\circ|Z,W;\xi)/\partial\xi|D^\circ;\theta\right\}\ , \end{aligned}$$

and

$$S_3(\zeta;\theta) = E_Z\left\{\partial \log P(Z|W;\zeta)/\partial\zeta|D^\circ;\theta\right\}\ .$$

An expectation-maximization algorithm (Dempster et al., 1977) can be carried out by solving the score equations iteratively so that if θ^{r-1} is the estimate at the $(r-1)$st iteration, the following system of equations is solved to obtain θ^r

$$\begin{aligned} S_1(\beta;\theta^{r-1}) &= E_Z\left\{\partial \log P(Y|Z,W;\beta)/\partial\beta|D^\circ;\theta^{r-1}\right\} = 0\ , \\ S_2(\xi;\theta^{r-1}) &= E_Z\left\{\partial \log P(X^\circ|Z,W;\xi)/\partial\xi|D^\circ;\theta^{r-1}\right\} = 0\ , \\ S_3(\zeta;\theta^{r-1}) &= E_Z\left\{\partial \log P(Z|W;\zeta)/\partial\zeta|D^\circ;\theta^{r-1}\right\} = 0\ , \end{aligned}$$

where the sum of contributions from individuals is taken to be implicit here. These expectations require computation of $P(Z|Y,X^\circ,W;\theta)$, which is given by

$$P(Z|Y,X^\circ,W;\theta) = \frac{P(Y|Z,W;\beta)P(X^\circ|Z,W;\xi)P(Z|W;\zeta)}{E_Z\{P(Y|Z,W;\beta)|X^\circ,W;\psi\}P(X^\circ|W;\psi)}\ ,$$

where $\psi = (\xi',\zeta')'$.

When all models are in the exponential family existing software can be used to solve the observed data score equation with respect to θ to obtain θ^r. We repeat this iterative procedure until a pre-specified convergence criterion is met, say, $\|\hat\theta^{r+1} - \hat\theta^r\| \le \epsilon$, where ϵ is a pre-defined tolerance. The value at the last iteration is the maximum likelihood estimator (MLE) $\hat\theta$. The variance of $\hat\theta$ can be estimated using the approach of Louis (1982) which allows us to extract the observed information matrix from functions related to the complete data log-likelihood employed in the EM algorithm. To be more specific, we let $\mathcal{S}(\theta)$ and $\mathcal{I}(\theta)$ denote the complete data score vector and information matrix respectively based on the log-likelihood function ℓ in (4). We then note that since $I(\theta) = E\left\{\mathcal{I}(\theta)|D^\circ\right\} - E\left\{\mathcal{S}(\theta)\mathcal{S}(\theta)'|D^\circ\right\}$, we can estimate the covariance of $\hat\theta$ based on the observed data information matrix $I(\theta)$ upon inserting the maximum likelihood estimate $\hat\theta$.

2.2 The Model for the Surrogate Variables

We now discuss the process leading to the availability of surrogate measurements in more detail. We suppose that the decision to observe more components of X depends on the observed components to that point. This corresponds to the situation in which an individual may receive imperfect diagnostic tests in an effort to determine their likely underlying state Z; investigators may choose to order more imperfect tests depending on the outcome of previous tests but this can be done in an *ad hoc* manner warranting a probability model for the process.

Under Assumption A1, $P(R|Y, Z, X^\circ, W) = P(R|Y, X^\circ, W)$ but further simplifications may be warranted. Since the decision on how many components of X should be measured is made at study entry, the following stronger assumption is often reasonable.

Assumption A3: $R \perp Y | X^\circ, W$.

Under Assumptions A1-A3, (1) can be rewritten as

$$\begin{aligned} L &\propto \sum_z P(R|X^\circ, W) P(Y|Z = z, W) P(X^\circ|Z = z, W) P(Z = z|W) \\ &= \sum_z P(Y|Z = z, W) P(R, X^\circ|Z = z, W) P(Z = z|W) \ . \end{aligned}$$

Note that we have re-introduced consideration of the missing data vector R to outline further assumptions and give a careful presentation of the assumptions for the missing data process for the surrogate variables. We assume that the surrogate variables are ordered corresponding to, for example, a sequence of diagnostic tests that may be requested by a treating physician. In particular, to reflect the sequential nature of the testing, if we let $\bar{R}_k = (R_1, \ldots, R_k)'$ and $\bar{X}_k = (X_1, \ldots, X_k)'$, we assume that the decision to order the kth test is based on a model of the form $P(R_k = r_k|\bar{R}_{k-1} = 1_{k-1}, \bar{X}_{k-1}, W)$ where $P(R_1 = 1) = 1$ and 1_{k-1} is a $(k-1) \times 1$ vector of ones. We may then write $P(R, X^\circ|Z, W)$ as

$$P(X_1|R_1 = 1, Z, W) \prod_{k=2}^{K} \left[P(X_k|\bar{R}_k = 1_k, \bar{X}_{k-1}, Z, W)^{R_k} P(R_k|\bar{R}_{k-1} = 1_{k-1}, \bar{X}_{k-1}, Z, W) \right]^{R_{k-1}}$$

from which we can focus on the partial likelihood

$$L(X^\circ, Z|R, W) \propto \prod_{k=2}^{K} P(X_k|\bar{R}_k = 1_k, \bar{X}_{k-1}, Z, W; \xi)^{R_k} P(X_1|R_1 = 1, Z, W; \xi) P(Z|W; \zeta) \ .$$

We may then write the partial likelihood as

$$L \propto \sum_z P(Y|Z = z, W) L(X^\circ, Z = z|R, W) \ . \tag{5}$$

We explore the use of this likelihood in the simulation studies of Section 3 and the application in Section 4. We first briefly discuss an alternative two-stage estimation procedure.

2.3 An Alternative Two-Stage Estimation Procedure

Note that we can also consider a two-stage estimation procedure in which the data from the surrogate variables are used alone for estimation of $\psi = (\xi', \zeta')'$. For the measurements made on the elements of X, we adopt a further conditional independence assumption.

Assumption A4: $P(X_k|\bar{R}_k = 1_k, \bar{X}_{k-1}, Z, W) = P(X_k|Z, W)$.

This states that the latent state determines the marginal distribution of each element of X, that there is conditional independence in the elements of X given Z, and that the availability of the preceding measures $\bar{X}_{k-1}$ does not affect the conditional probability $P(X_k|Z,W)$.

By Assumption A4, we can write

$$L(X^\circ|R,W;\psi) \propto \sum_z \prod_{k=2}^{K} P(X_k|Z,W;\xi)^{R_k} P(X_1|Z,W;\xi) P(Z|W;\zeta) \; . \tag{6}$$

This partial likelihood does not attempt to exploit information from the response when estimating ψ where $\psi = (\xi', \zeta')'$.

As pointed out by Walter and Irwig (1988) and Liu and Liang (1991), at least three surrogate variables are required for some individuals for the parameters to be identifiable under the conditional independence Assumption A4. In such a setting, an EM algorithm based on (6) is straightforward to implement to obtain $\tilde{\psi}$. That is, one could focus on the complete data log-likelihood corresponding to $L(X^\circ|R,W)$ in (6) as

$$\sum_z I(Z=z) \left[\sum_{k=2}^{K} R_k \log P(X_k|Z,W) + \log P(X_1|Z,W) + \log P(Z|W) \right]$$

and iterate following the EM algorithm to maximize $L(X^\circ|R,W)$ with respect to ψ and obtain its MLE $\tilde{\psi}$. The E-step requires computation of $P(Z|X^\circ,W;\psi)$ which is given by

$$P(Z|X^\circ,W;\psi) = \frac{P(X^\circ|Z,W;\xi)P(Z|W;\zeta)}{\sum_z P(X^\circ|Z=z,W;\xi)P(Z=z|W;\zeta)} \; .$$

Again one can use the approach of Louis (1982) to get a covariance matrix for $\tilde{\psi}$.

The estimate $\tilde{\psi}$ can then be inserted in (5) to give $L(\beta,\tilde{\psi})$ which may be maximized with respect to β to give a two-stage estimator $\tilde{\beta}$; variance estimation can be carried out as in Shih and Louis (1995). Alternatively an EM algorithm could also be used for estimation to obtain $\tilde{\beta}$ by inserting $\tilde{\psi}$ into (5) and using the corresponding complete data likelihood to iteratively update β^r by applying this to $S_1(\beta;\beta^{r-1},\tilde{\psi})$, where the E-step involves the computation of

$$\frac{P(Y|Z,W;\beta)L(X^\circ,Z|R,W;\tilde{\psi})}{\sum_z P(Y|Z=z,W;\beta)L(X^\circ,Z=z|R,W;\tilde{\psi})} \; .$$

There would be efficiency and robustness tradeoffs with this approach, and we have focussed on maximum likelihood estimation in this manuscript for reasons of efficiency.

3 Simulation Studies

Not all individuals in a study will be measured by a minimum of $K=3$ instruments. That is, $K_i \le K$ for some $i = 1,\ldots,n$. It is easy to see that there are in total seven possible missing patterns for the three surrogates of Z_i. Here, we focus on a general scenario that often arises in clinical studies representing a sequential missing pattern as in the stimulating study on breast cancer to be discussed in Section 4. That is, response Y_i, precisely observed covariate vector W_i and measurement by first instrument X_{i1} are observed for all subjects, $i=1,\ldots,n$. However, the probability of being assessed by the second instrument and observing X_{i2} depends on the observed data X_{i1} and W_i. If a subject is not assessed by the second instrument, they will not be measured by the third instrument. If a subject is measured by the second instrument, the probability of being assessed by the third instrument depends on X_{i1}, X_{i2} and W_i. This describes a missing at random mechanism for the surrogates (Rubin, 1976). Here, a subject may have a higher probability of being measured by the other instruments when the subject is measured positive or negative according to the previous instruments depending on the particular study design and clinical practice. It results in a biased sub-sample if we restrict attention to individuals measured by all instruments.

3.1 Simulation Studies Involving Complete Surrogate Data

Here, we conduct simulation studies to evaluate the empirical performances of the proposed method for fitting logistic regression models involving misclassified covariates in the absence of validation data where the surrogates are all available. Without losing generality, here we do not consider an auxiliary variable W_i, but focus on a regression model for $Y|Z$. For subject i, we first generate the binary latent variable Z_i as a Bernoulli random variable with $P(Z_i = 1)$ = $\exp(\zeta)/(1 + \exp(\zeta))$=0.25,0.50, 0.75. We then generate the binary response Y_i with conditional probability $P(Y_i = 1|Z_i = z_i)$ = $\exp(\beta_0 + \beta_1 z_i)/(1 + \exp(\beta_0 + \beta_1 z_i))$. We set $\beta_1 = \log 1.5$ so that the odds that the response is 1 is 50% higher when $Z_i = 1$ vs $Z_i = 0$. We determine the value of β_0 to ensure $P(Y_i) = 0.25, 0.50, 0.75$. Let $K = 3$, we generate three conditionally independent surrogate values X_{ik}, $k = 1, 2, 3$, given Z_i with specified sensitivity $P(X_{ik} = 1|Z_i = 1) = \alpha_k$ and specificity $P(X_{ik} = 0|Z_i = 0) = \gamma_k$ which can be parameterized by vector $\xi_k = (\xi_{k0}, \xi_{k1})'$ through another logistic regression model $\text{logit}P(X_{ik} = 1|Z_i = z_i; \xi_k) = \xi_{k0} + \xi_{k1} z_i$; we let $\xi = (\xi_1', \xi_2', \xi_3')'$. We investigate performance with different misclassification rates for the surrogate variables. For simplicity, we set the same sensitivity and specificity for all three surrogate measures ($\alpha_k = \alpha$, $\gamma_k = \gamma$), $k = 1, 2, 3$ and set α=γ=0.95,0.85,0.75.

Table 1: Empirical bias (EBIAS), empirical standard error (ESE), average model-based standard error (ASE) and empirical coverage probability (ECP) for the estimates of β and ζ based on different analysis with complete data on the surrogate variables when $P(Y = 1) = 0.5$, $P(Z = 1) = 0.5$, $n = 1000$, $nsim = 500$.

		$\beta_0 = -0.203$				$\beta_1 = 0.405$				$\zeta = 0$			
Method	Detail	EBIAS	ESE	ASE	ECP	EBIAS	ESE	ASE	ECP	EBIAS	ESE	ASE	ECP
						sensitivity = specificity = 0.95							
ORACLE	True Z	0.000	0.086	0.090	95.4	-0.003	0.125	0.127	95.2	-0.000	0.061	0.063	96.4
EM	Full	-0.000	0.087	0.091	96.2	-0.002	0.127	0.129	95.6	-0.001	0.063	0.065	96.2
EM	Two-Stage	0.005	0.086	0.090	95.8	-0.013	0.124	0.127	95.4	-0.001	0.063	0.065	96.2
Naive	X_1	0.018	0.085	0.090	96.8	-0.040	0.123	0.127	93.8	-0.001	0.059	0.063	95.4
Ad Hoc	$Z^* = I(X_\cdot \geq 1)$	-0.000	0.092	0.097	96.6	-0.053	0.125	0.129	93.8	0.287	0.062	0.064	0.4
	$Z^* = I(X_\cdot \geq 2)$	0.003	0.087	0.090	95.8	-0.008	0.126	0.127	95.0	-0.001	0.061	0.063	96.2
	$Z^* = I(X_\cdot = 3)$	0.049	0.080	0.084	92.0	-0.051	0.128	0.129	94.6	-0.287	0.062	0.064	0.6
						sensitivity = specificity = 0.85							
ORACLE	True Z	0.000	0.086	0.090	95.4	-0.003	0.125	0.127	95.2	-0.000	0.061	0.063	96.4
EM	Two-Stage	0.041	0.083	0.090	95.0	-0.085	0.114	0.127	92.0	0.001	0.096	0.092	94.2
EM	Full	-0.001	0.093	0.096	96.0	-0.000	0.144	0.143	94.6	0.000	0.095	0.092	94.6
Naive	X_1	0.060	0.087	0.090	91.4	-0.123	0.128	0.127	83.0	0.001	0.059	0.063	96.4
Ad Hoc	$Z^* = I(X_\cdot \geq 1)$	0.002	0.108	0.115	97.4	-0.118	0.133	0.138	85.8	0.807	0.065	0.068	0.0
	$Z^* = I(X_\cdot \geq 2)$	0.023	0.088	0.090	95.4	-0.049	0.128	0.127	92.6	-0.000	0.063	0.063	95.2
	$Z^* = I(X_\cdot = 3)$	0.113	0.075	0.076	69.2	-0.119	0.139	0.138	85.2	-0.808	0.067	0.069	0.0

Table 2: Empirical bias (EBIAS), empirical standard error (ESE), average model-based standard error (ASE) and empirical coverage probability (ECP) for the estimates of ξ based on different analysis with complete data on the surrogate variables when $P(Y=1)=0.5$, $P(Z=1)=0.5$, $n=1000$, $nsim=500$.

			sensitivity = specificity = 0.95								sensitivity = specificity = 0.85							
			$\xi_0=-2.944$				$\xi_1=5.889$				$\xi_0=-1.735$				$\xi_1=3.469$			
Surrogate	Method	Detail	EBIAS	ESE	ASE	ECP	EBIAS	ESE	ASE	ECP	EBIAS	ESE	ASE	ECP	EBIAS	ESE	ASE	ECP
X_1	ORACLE	True Z	-0.01	0.20	0.21	95.2	0.02	0.29	0.29	95.8	0.00	0.11	0.13	97.0	0.00	0.17	0.18	96.4
	EM	Full	-0.01	0.22	0.23	96.6	0.03	0.31	0.32	96.4	-0.01	0.17	0.18	96.6	0.03	0.23	0.24	96.6
	EM	Two-Stage	-0.01	0.22	0.23	96.6	0.03	0.31	0.32	96.2	-0.01	0.17	0.18	97.0	0.03	0.23	0.24	96.2
	Naive	X_1																
	Ad Hoc	$Z^*=I(X.\geq 1)$	-17.44	0.38	795.52	100.0	16.45	0.46	795.52	100.0	-17.80	0.19	603.44	100.0	17.03	0.21	603.44	100.0
		$Z^*=I(X.\geq 2)$	-0.07	0.20	0.21	95.6	0.13	0.29	0.30	94.8	-0.19	0.13	0.13	72.6	0.39	0.18	0.19	45.4
		$Z^*=I(X.=3)$	0.99	0.12	0.13	0.0	16.49	0.43	805.04	100.0	0.78	0.08	0.09	0.0	17.00	0.26	598.11	100.0
X_2	ORACLE	True Z	-0.02	0.21	0.21	96.4	0.06	0.31	0.30	94.2	-0.00	0.12	0.13	96.2	0.00	0.18	0.18	94.4
	EM	Full	-0.02	0.23	0.23	94.4	0.08	0.35	0.33	93.4	-0.01	0.17	0.18	95.4	0.02	0.24	0.24	95.0
	EM	Two-Stage	-0.02	0.23	0.23	94.6	0.08	0.35	0.33	93.2	-0.01	0.17	0.18	95.6	0.02	0.24	0.24	94.8
	Naive	X_1	0.68	0.15	0.15	2.2	-1.35	0.22	0.22	0.0	0.66	0.09	0.10	0.0	-1.32	0.14	0.15	0.0
	Ad Hoc	$Z^*=I(X.\geq 1)$	-17.46	0.37	801.02	100.0	16.48	0.44	801.02	100.0	-17.78	0.23	598.44	100.0	17.00	0.26	598.44	100.0
		$Z^*=I(X.\geq 2)$	-0.07	0.22	0.21	94.4	0.18	0.32	0.31	91.4	0.77	0.08	0.09	0.0	17.01	0.23	601.12	100.0
		$Z^*=I(X.=3)$	1.00	0.13	0.13	0.0	16.47	0.43	802.25	100.0	-0.19	0.13	0.13	74.8	0.38	0.19	0.19	51.0
X_3	ORACLE	True Z	-0.03	0.20	0.21	96.2	0.04	0.29	0.29	95.4	-0.01	0.12	0.13	96.0	0.02	0.17	0.18	97.0
	EM	Full	-0.02	0.22	0.23	96.2	0.04	0.31	0.32	95.0	-0.02	0.18	0.18	96.6	0.03	0.24	0.24	95.8
	EM	Two-Stage	-0.02	0.22	0.23	96.2	0.04	0.31	0.32	95.0	-0.02	0.18	0.18	96.2	0.03	0.24	0.24	95.8
	Naive	X_1	0.68	0.15	0.15	2.2	-1.37	0.20	0.22	0.0	0.66	0.10	0.10	0.0	-1.32	0.14	0.15	0.0
	Ad Hoc	$Z^*=I(X.\geq 1)$	-17.46	0.37	800.25	100.0	16.46	0.44	800.25	100.0	-17.78	0.21	600.42	100.0	17.01	0.24	600.42	100.0
		$Z^*=I(X.\geq 2)$	-0.07	0.21	0.21	96.8	0.14	0.30	0.30	95.2	-0.20	0.13	0.14	68.6	0.39	0.19	0.19	47.8
		$Z^*=I(X.=3)$	0.98	0.13	0.13	0.0	16.49	0.44	803.81	100.0	0.77	0.09	0.09	0.0	17.01	0.26	599.10	100.0

Five hundred datasets of size 500 or 1000 were simulated for each parameter configuration. Here, the naive methods refer to the scenario that we simply ignore the issue of misclassification in the surrogates and take X_{ik} as the true latent variable Z_i, $k = 1, 2, 3$, or naively consider them jointly to reach a consensus to create a new *ad hoc* and fit the regression models directly. For example, researchers may naively take the unknown Z_i as positive if at least k of the three surrogate values are positive, and otherwise negative, denoted by $Z_i^* = I(X_{i\cdot} \geq k)$, $k = 1, 2, 3$, where $X_{i\cdot} = \sum_{k=1}^{3} X_{ik}$. The proposed method in Section 2 is denoted by EM-Full, where the standard errors (SEs) are obtained using Louis's method (Louis, 1982). The performances of the different estimators are summarized in Tables 1 to 2 in terms of the empirical bias (EBIAS), empirical standard error (ESE), average model-based standard error (ASE) and empirical coverage probability (ECP), defined as the fraction of simulations for which the sample confidence interval (CI) contains the true parameter value. Here, Table 1 summarizes the simulation results for the parameter vector estimation corresponding to the regression model and latent variable distribution. Table 2 is for the accuracy of the instruments.

If the instruments are relatively accurate, say, with 0.95 sensitivity and specificity, as shown in the upper panel of Table 1, naively using one of them as the latent variable may provide only slighted biased estimates of β and ζ indexing the regression coefficient and latent variable distribution, respectively, but the estimates of ζ representing the evaluation of the instruments are severely biased as shown in the left panel of Table 2. Given that we set the accuracy of all three participating instruments the same and here we consider the case that all the surrogate values are available, only results when naively treating X_1 as Z are included in the tables. When the instruments are less accurate, say, with 0.85 sensitivity and specificity as noted in the lower panel of Table 1 and the right panel of Table 2, the estimates of β and ζ are more appreciably biased. Either way, naively assembling the surrogate values accordingly to certain rules to reach a consensus and serve as the latent variable, generally produces biased estimates of the parameters of interest especially when the accuracy of the instruments are lower based on the results in Tables 1 and 2. The proposed full EM algorithm, is relatively robust to the accuracy of the instruments. It gives comparable results to the analysis based on the true latent variable with slightly larger standard errors. The empirical biases are generally small though the standard errors increase slightly when the accuracy of the instruments decreases, and there is good agreement between the empirical and average model-based standard errors and the empirical coverage probabilities are compatible with the nominal 95% level.

3.2 Simulation Studies Involving Incomplete Surrogate Data

When simulating the missingness in the surrogate values, we mimic the sequential missing pattern as discussed in the motivating study; see Section 4. That is, in a sample with n independent individuals, we have three subsets

$$\begin{aligned}
\mathcal{S}_1 &= \{i : D_{3i} = (Y_i, X_i^\circ = X_{i1}, R_i = (1,0,0)')'\} \ , \\
\mathcal{S}_2 &= \{i : D_{2i} = (Y_i, X_i^\circ = (X_{i1}, X_{i2})', R_i = (1,1,0)')'\} \ , \\
\mathcal{S}_3 &= \{i : D_{1i} = (Y_i, X_i^\circ = (X_{i1}, X_{i2}, X_{i3})', R_i = (1,1,1)')'\} \ ,
\end{aligned}$$

with size n_1, n_2 and n_3 and proportion p_1, p_2 and p_3 respectively, where $p_j = n_j/n$, $j = 1, 2, 3$. To be specific, we let $P(R_{i1} = 1) = 1$, and logit $P(R_{i2} = 1|R_{i1} = 1, X_{i1} = x_{i1}; a) = a_0 + a_1 x_{i1}$ where $a = (a_0, a_1)'$, $a_1 = \log 1.50$ so that when X_{i1} is observed, the odds of having X_{i2} observed are 50% higher in the group where $X_{i1} = 1$. We can determine a_0 so that the proportion of subset $\mathcal{S}_3$ and $\mathcal{S}_2$ is $p_3 + p_2$. Similarly, we set $P(R_{i3} = 0|R_{i1} = 1, R_{i2} = 0) = 1$, and logit $P(R_{i3} = 1|R_{i1} = 1, R_{i2} = 1, X_{i1}, X_{i2}; b) = b_0 + b_1 x_{i1} + b_2 x_{i2}$, where $b = (b_0, b_1, b_2)'$, $b_1 = b_2 = \log 1.25$ so that when X_{i1} and X_{i2} are both observed, the odds of having X_{i3} observed are higher when at least one of the first two instruments test positive. We solve for b_0 to ensure the proportion of subset $\mathcal{S}_3$ is p_3, i.e., $P(R_{i3} = 1|R_{i2} = 1, R_{i1} = 1) = p_3/(p_3 + p_2)$. We can set (p_3, p_2, p_1) to different values to represent different amounts of missingness of the surrogates, say $(0.50, 0.25, 0.25)$, $(0.40, 0.30, 0.30)$ and $(0.30, 0.35, 0.35)$, or $(1, 0, 0)$ to represent the case that there are no missingness in surrogates.

Table 3: Empirical bias (EBIAS), empirical standard error (ESE), average model-based standard error (ASE) and empirical coverage probability (ECP) for the estimates of β and ζ based on different analysis with incomplete surrogate data when $P(Y = 1) = 0.5$, $P(Z = 1) = 0.5$, sensitivity=specificity=0.95, $n = 1000$, $nsim = 500$.

		$\beta_0 = -0.203$				$\beta_1 = 0.405$				$\zeta = 0$			
Method	Detail	EBIAS	ESE	ASE	ECP	EBIAS	ESE	ASE	ECP	EBIAS	ESE	ASE	ECP
$(p_3, p_2, p_1) = (0.5, 0.25, 0.25)$													
Complete Case	True Z	0.005	0.135	0.135	95.2	-0.008	0.181	0.181	95.4	0.226	0.086	0.090	27.4
Complete Case	EM	0.004	0.136	0.136	94.8	-0.007	0.184	0.184	95.2	0.224	0.090	0.092	30.2
Full Data	EM	-0.000	0.088	0.092	96.4	-0.002	0.132	0.133	96.4	-0.002	0.071	0.070	95.0
$(p_3, p_2, p_1) = (0.4, 0.3, 0.3)$													
Complete Case	True Z	0.003	0.150	0.154	96.6	-0.007	0.198	0.204	95.6	0.282	0.101	0.101	20.8
Complete Case	EM	0.002	0.150	0.155	97.2	-0.004	0.200	0.207	96.0	0.279	0.105	0.103	21.8
Full Data	EM	-0.000	0.088	0.093	96.0	-0.002	0.132	0.134	95.8	-0.002	0.072	0.072	95.0
$(p_3, p_2, p_1) = (0.3, 0.35, 0.35)$													
Complete Case	True Z	0.005	0.180	0.181	94.6	-0.001	0.240	0.236	95.4	0.338	0.116	0.117	17.4
Complete Case	EM	0.005	0.183	0.182	94.8	-0.000	0.245	0.240	95.4	0.336	0.121	0.120	19.2
Full Data	EM	-0.001	0.087	0.093	96.6	-0.001	0.129	0.135	97.2	0.000	0.073	0.076	95.8

Table 4: Empirical bias (EBIAS), empirical standard error (ESE), average model-based standard error (ASE) and empirical coverage probability (ECP) for the estimates of β and ζ based on different analysis with incomplete surrogate data when $P(Y = 1) = 0.5$, $P(Z = 1) = 0.5$, sensitivity=specificity=0.85, $n = 1000$, $nsim = 500$.

		$\beta_0 = -0.203$				$\beta_1 = 0.405$				$\zeta = 0$			
Method	Detail	EBIAS	ESE	ASE	ECP	EBIAS	ESE	ASE	ECP	EBIAS	ESE	ASE	ECP
$(p_3, p_2, p_1) = (0.5, 0.25, 0.25)$													
Complete Case	True Z	0.003	0.133	0.133	96.2	-0.008	0.179	0.181	95.4	0.175	0.086	0.090	48.2
Complete Case	EM	0.004	0.147	0.144	95.0	-0.010	0.211	0.205	95.2	0.172	0.137	0.132	74.6
Full Data	EM	0.001	0.096	0.101	97.4	-0.003	0.156	0.155	95.0	-0.004	0.128	0.120	94.4
$(p_3, p_2, p_1) = (0.4, 0.3, 0.3)$													
Complete Case	True Z	0.003	0.147	0.151	96.4	-0.010	0.199	0.203	95.8	0.219	0.100	0.101	40.6
Complete Case	EM	0.002	0.160	0.163	95.6	-0.005	0.230	0.230	95.6	0.214	0.156	0.148	68.2
Full Data	EM	0.001	0.096	0.102	96.8	-0.004	0.157	0.157	94.8	-0.003	0.138	0.132	94.2
$(p_3, p_2, p_1) = (0.3, 0.35, 0.35)$													
Complete Case	True Z	0.007	0.176	0.177	95.6	-0.010	0.240	0.235	95.4	0.265	0.116	0.117	37.8
Complete Case	EM	0.006	0.200	0.192	93.4	-0.007	0.285	0.267	93.4	0.264	0.177	0.173	67.6
Full Data	EM	0.000	0.099	0.104	97.2	-0.005	0.160	0.160	94.6	0.006	0.144	0.149	95.0

When surrogates are missing with a sequential pattern, different proportions of $\mathcal{S}_1$, $\mathcal{S}_2$ and $\mathcal{S}_3$ are set as in Tables 3, 4 and 5. The analysis based on the individuals with complete information of all surrogates and the true latent variable information, confirms that the subset of complete cases

Table 5: Empirical bias (EBIAS), empirical standard error (ESE), average model-based standard error (ASE) and empirical coverage probability (ECP) for the estimates of ξ based on different analysis with incomplete surrogate data when $P(Y = 1) = 0.5$, $P(Z = 1) = 0.5$, $n = 1000$, $nsim = 500$.

			sensitivity = specificity = 0.95								sensitivity = specificity = 0.85							
			$\xi_0 = -2.944$				$\xi_1 = 5.889$				$\xi_0 = -1.735$				$\xi_1 = 3.469$			
Surrogate	Method	Detail	EBIAS	ESE	ASE	ECP	EBIAS	ESE	ASE	ECP	EBIAS	ESE	ASE	ECP	EBIAS	ESE	ASE	ECP
									$(p_3, p_2, p_1) = (0.5, 0.25, 0.25)$									
X_1	Complete Case	True Z	0.16	0.31	0.29	88.8	0.04	0.43	0.42	95.2	0.18	0.17	0.18	80.6	0.00	0.25	0.25	95.2
	Complete Case	EM	0.16	0.34	0.32	87.6	0.05	0.48	0.47	94.8	0.16	0.25	0.25	84.2	0.05	0.36	0.35	97.6
	Full Data	EM	-0.03	0.31	0.29	94.6	0.07	0.44	0.43	96.0	-0.02	0.23	0.23	95.0	0.06	0.33	0.33	96.8
X_2	Complete Case	True Z	0.05	0.30	0.30	93.6	0.06	0.41	0.42	96.0	0.08	0.18	0.18	92.6	0.00	0.26	0.25	95.8
	Complete Case	EM	0.04	0.35	0.34	93.0	0.10	0.48	0.47	95.2	0.07	0.27	0.26	90.4	0.03	0.36	0.35	95.6
	Full Data	EM	-0.04	0.32	0.31	95.8	0.11	0.44	0.43	96.8	-0.02	0.25	0.24	94.4	0.04	0.33	0.32	94.4
X_3	Complete Case	True Z	-0.05	0.33	0.32	96.4	0.10	0.44	0.43	94.8	-0.02	0.20	0.19	95.0	0.02	0.26	0.25	94.8
	Complete Case	EM	-0.05	0.39	0.36	96.8	0.11	0.51	0.48	94.8	-0.04	0.30	0.27	95.6	0.07	0.36	0.36	95.6
	Full Data	EM	-0.05	0.39	0.36	96.8	0.11	0.51	0.47	94.2	-0.03	0.29	0.27	95.6	0.06	0.35	0.35	95.4
									$(p_3, p_2, p_1) = (0.4, 0.3, 0.3)$									
X_1	Complete Case	True Z	0.18	0.35	0.33	86.2	0.06	0.50	0.48	94.8	0.22	0.19	0.20	77.8	0.01	0.28	0.28	95.4
	Complete Case	EM	0.18	0.40	0.36	87.6	0.08	0.56	0.53	94.8	0.20	0.29	0.28	83.8	0.07	0.42	0.41	97.0
	Full Data	EM	-0.04	0.35	0.32	96.2	0.08	0.50	0.48	96.4	-0.03	0.27	0.25	96.0	0.08	0.38	0.37	96.2
X_2	Complete Case	True Z	0.06	0.33	0.35	94.0	0.10	0.48	0.48	96.0	0.09	0.20	0.20	91.6	0.01	0.29	0.28	95.8
	Complete Case	EM	0.05	0.39	0.39	92.0	0.14	0.56	0.54	95.6	0.08	0.31	0.30	91.0	0.04	0.41	0.40	96.2
	Full Data	EM	-0.04	0.36	0.34	96.0	0.13	0.49	0.48	96.8	-0.02	0.27	0.27	94.6	0.05	0.35	0.36	97.2
X_3	Complete Case	True Z	-0.08	0.40	0.37	95.8	0.13	0.52	0.49	95.4	-0.03	0.23	0.21	94.8	0.04	0.30	0.29	94.8
	Complete Case	EM	-0.08	0.46	0.42	96.4	0.15	0.59	0.55	96.0	-0.06	0.34	0.32	95.0	0.10	0.42	0.41	96.0
	Full Data	EM	-0.08	0.45	0.41	96.6	0.15	0.58	0.55	95.8	-0.05	0.34	0.31	94.4	0.09	0.41	0.40	95.8
									$(p_3, p_2, p_1) = (0.3, 0.35, 0.35)$									
X_1	Complete Case	True Z	0.19	0.39	0.38	87.2	0.11	0.57	0.57	96.8	0.23	0.23	0.23	79.4	0.03	0.32	0.33	96.4
	Complete Case	EM	0.18	0.46	0.43	88.0	0.14	0.64	0.64	97.0	0.22	0.34	0.33	84.2	0.07	0.45	0.47	97.4
	Full Data	EM	-0.06	0.38	0.37	97.6	0.11	0.53	0.55	98.0	-0.05	0.29	0.29	97.4	0.09	0.41	0.42	96.0
X_2	Complete Case	True Z	0.04	0.43	0.41	92.8	0.17	0.58	0.57	96.4	0.10	0.23	0.24	92.6	0.02	0.33	0.33	95.4
	Complete Case	EM	0.03	0.51	0.47	92.2	0.20	0.69	0.65	96.6	0.09	0.34	0.35	93.2	0.05	0.46	0.47	97.0
	Full Data	EM	-0.07	0.43	0.41	96.6	0.17	0.59	0.57	96.8	-0.04	0.30	0.31	97.0	0.07	0.40	0.41	96.6
X_3	Complete Case	True Z	-0.08	0.47	0.44	95.6	0.14	0.60	0.57	95.8	-0.01	0.24	0.25	95.2	0.03	0.31	0.33	96.2
	Complete Case	EM	-0.10	0.53	0.51	96.2	0.19	0.68	0.66	97.2	-0.07	0.42	0.40	96.6	0.12	0.50	0.50	97.0
	Full Data	EM	-0.10	0.53	0.51	95.6	0.19	0.68	0.66	97.2	-0.06	0.47	0.57	97.0	0.10	0.54	0.67	97.0

represents a biased sub-sample. This is reflected by the biases in the parameter vector indexing the latent variable distribution ζ and assessment of the instrument accuracy ξ. Here, higher accuracy of the instruments and/or bigger missing proportions lead to bigger biases in the estimates of ζ as shown in Tables 3 and 4, which is clear to see in Figure 1. Similarly, the proposed EM algorithm based on the complete cases results in biased estimates of ζ and ξ. The biases for ζ are noticeably larger with bigger missingness proportion when the accuracy of the instruments is fixed, so are they if the identification tools are more accurate when the missing percentage is held the same. The proposed method applied on the full data including individuals with missing surrogates, successfully resolves the issue. It produces generally small empirical biases, closely agreed empirical and average standard errors and empirical coverage probabilities around the nominal level of 95% though the associate standard errors increase accordingly when the proportion of missingness increases with the same sets of instruments. The bigger bias in ζ based on complete cases when the instruments are more accurate is successfully addressed in the proposed method dealing with missingness in surrogates. Figure 1 gives clear visual display of the comparisons and good performance of the proposed method on the estimation of the latent variable distribution in the absence of missingness in surrogates and in the presence of it.

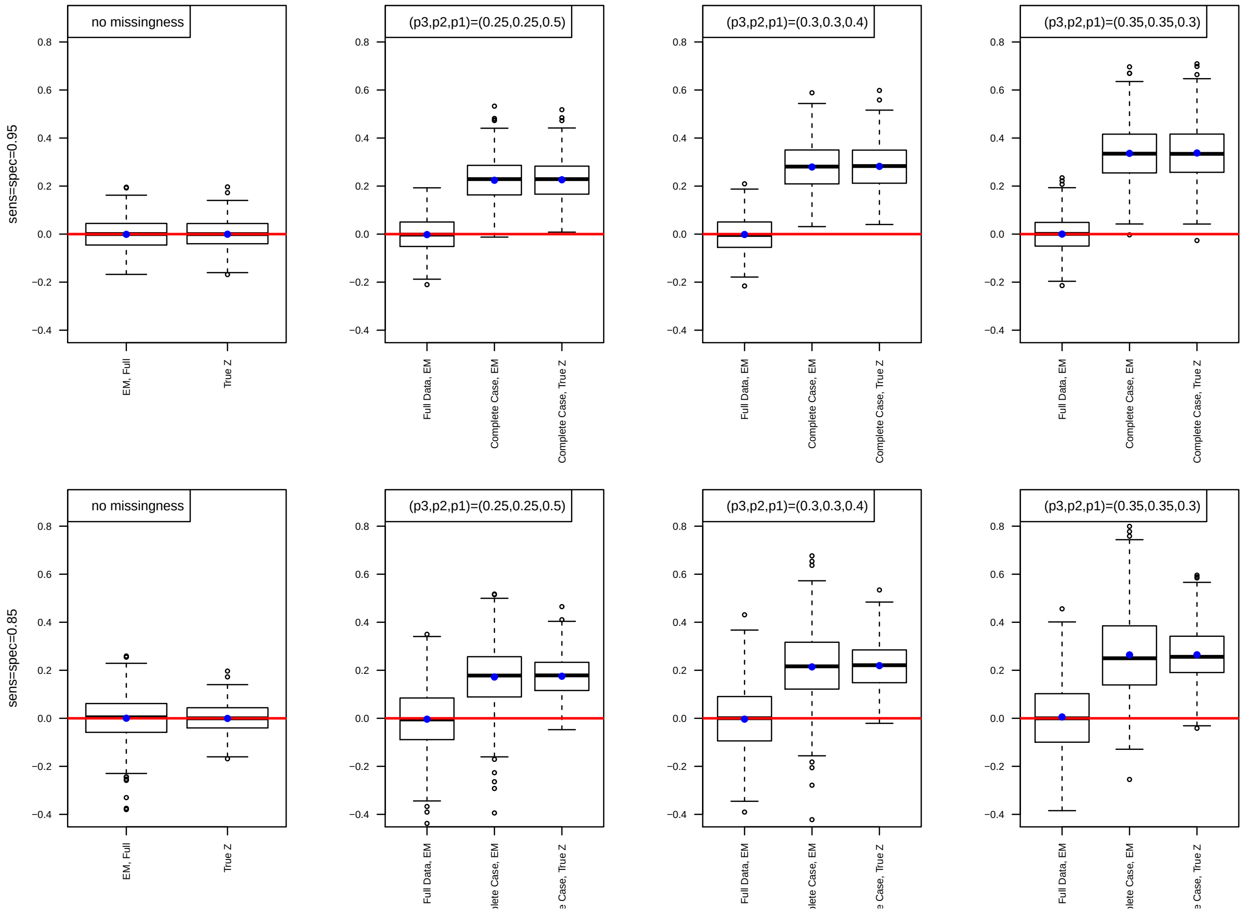

Figure 1: Box plots of the estimates of ζ corresponding to the distribution of the latent variable from different analysis in the absence or presence of missing data when $P(Y=1)=0.5$, $P(Z=1)=0.5$, sensitivity=specificity=0.95 (upper panel) or 0.85 (lower panel), $n = 1000$, $nsim = 500$. Here, the horizontal red lines running across the sub-figures correspond to true value ζ and solid blue dots show the averages of the estimates.

4 Application to an Observational Breast Cancer Study

A study of patients with breast cancer was carried out at the Tom Baker Cancer Centre, Calgary, Canada (Feng et al., 2015, 2016). A sub-analysis involves a group of 235 individuals who were diagnosed with early-stage breast cancer (stage I, II or III), had no chemotherapy and were treated with Tamoxifen for hormone-receptor positive breast cancer (estrogen receptor positive or progesterone receptor positive) with negative human epidermal growth factor receptor 2 (HER2) status. The goal is to investigate the association between potential risk factors and clinical outcomes in the subgroup. In particular, researchers are interested in knowing whether the expression of a particular serine/threonine protein kinase ataxia telangiectasia mutated (ATM) in both tumor and cancer-associated stromal is significant when adjusting for other prognostic factors including tumor grade, tumor size, lymph node (LN) status, lymphovascular invasion (LVI) and age at cancer diagnosis. The status outcomes of interest is whether patients die of breast cancer within five years of diagnosis. The distribution of patients' characteristics with respect to the outcome is summarized in Table 7. The protein expression of ATM in both malignant tumor and stromal compartments are measured using fluorescent immunohistochemistry and automated quantitative analysis (AQUA). The biomarkers are measured by different number of labs in Alberta Health Service: three times on 80 (34.0%) patients, twice on 80 (34.0%) patients and only once on 75 (31.9%) patients. The resulting continuous measurements are discrepant and subject to measurement error. An AQUA score of 72.2 is used to further dichotomize the patients into ATM high or lower groups. This cut-point is independently identified by X-Tile software to avoid potential human bias (Camp et al., 2004). However, misclassification and missingness in surrogates arises consequently: according the first lab, 123 (52.3%) are high ATM and 112 (47.7%) are low ATM; according the second lab, 80 (34.0%) are high ATM, 80 (34.0%) are low ATM and 75 (31.9%) are missing; according the third lab, 36 (15.3%) are high ATM, 44 (18.7%) are low ATM and 155 (66.0%) are missing. Table 6 further displays the joint results of the ATM assessments from three labs indicating the sequential missing data pattern discussed before. The methodology proposed in Section 2 is therefore suitable to address this challenge.

Tables 8 and 9 give the results of fitting multivariate models and applying different analysis on complete cases (CCs) only without missingness in surrogates and available cases (ACs) including subjects who are not measured by all three labs. The naive analysis considered include treating the kth misclassification as the true classification and the ad hoc approach takes it as positive when at least k surrogate values are positive otherwise negative, $k = 1, 2, 3$. We present the results based on complete cases following the proposed EM algorithm in Section 2. We also follow the procedure described in Section 2 on full data involving missingness in the surrogate values. The estimated odds ratio with respect to patient status (deceased vs alive) and corresponding 95% CI are presented in Table 8 where the significant effects are highlighted in boldface. We find that LN status is statistically significantly associated with the outcome when adjusting for other risk factors in all analysis. The effects of ATM and/or age at diagnosis are statistically significant in addition in some of the naive analysis involving complete case analysis or available case analysis. ATM is not significant after adjusting for all other five risk factors when applying the proposed method to complete case analysis or available case analysis. The estimated proportion of high ATM, sensitivity and specificity of each lab and the corresponding 95% CIs based on different approaches are available in Table 9. The proposed method based on available case analysis produces narrower 95% CIs than those on complete case analysis as shown in both tables. The point estimates of sensitivity of all three labs are high and those of specificity are moderate, and they are associated with large variability possibly due to small sample size and patient heterogeneity.

Table 6: A $2 \times 3 \times 3$ table on the joint results of the three labs providing some data on ATM assessments.

	Lab 1								
	Negative				Positive				
	Lab 2				Lab 2				
Lab 3	Negative	Positive	Missing	Total	Negative	Positive	Missing	Total	Total
Negative	34	3	0	37	5	2	0	7	44
Positive	6	3	0	9	3	24	0	27	36
Missing	28	9	29	66	4	39	46	89	155
Total	68	15	29	112	12	65	46	123	235

Table 7: Summary of patients' characteristics with respect to status within 5 years of cancer diagnosis due to breast cancer.

		Death		
Variable	Category	Yes	No	Overall
Tumor grade	1 or 2	180(86.1%)	16(61.5%)	196(83.4%)
	3	29(13.9%)	10(38.5%)	39(16.6%)
Tumor size	< 5cm	204(97.6%)	22(84.6%)	226(96.2%)
	≥ 5cm	5(2.4%)	4(15.4%)	9(3.8%)
LN	absent	166(79.4%)	10(38.5%)	176(74.9%)
	present	43(20.6%)	16(61.5%)	59(25.1%)
LVI	absent	173(82.8%)	12(46.2%)	185(78.7%)
	present	36(17.2%)	14(53.8%)	50(21.3%)
Age at diagnosis	< 65 years	110(52.6%)	5(19.2%)	115(48.9%)
	≥ 65 years	99(47.4%)	21(80.8%)	120(51.1%)
	Overall	209(88.9%)	26(11.1%)	235

Table 8: Estimated odds ratio and associated 95% CIs and the corresponding sample size (n) using different methods in the breast cancer study.

		OR and 95%CI						
Method	Detail	Z	W_1	W_2	W_3	W_4	W_5	n
					Complete Cases			
Naive	X_1	0.35(0.07,1.84)	0.48(0.09,2.70)	6.33(0.38,104.28)	2.25(0.48,10.42)	**6.26(1.32,29.66)**	6.25(0.97,40.24)	80
	X_2	0.28(0.05,1.61)	0.58(0.11,3.03)	6.51(0.41,102.17)	2.27(0.49,10.51)	**5.51(1.20,25.20)**	6.10(0.96,38.72)	80
	X_3	0.30(0.06,1.49)	0.47(0.08,2.58)	10.69(0.58,196.26)	2.52(0.54,11.77)	**5.54(1.19,25.72)**	5.16(0.84,31.66)	80
Ad Hoc	$Z^* = I(X. \geq 1)$	**0.13(0.03,0.67)**	0.41(0.07,2.41)	8.67(0.37,203.07)	2.01(0.41,9.80)	**6.51(1.24,34.14)**	**7.36(1.07,50.78)**	80
	$Z^* = I(X. \geq 2)$	0.51(0.10,2.60)	0.56(0.10,2.99)	7.86(0.48,127.74)	2.40(0.53,10.97)	**5.73(1.28,25.76)**	5.45(0.90,33.22)	80
	$Z^* = I(X. \geq 3)$	0.50(0.08,3.08)	0.57(0.11,3.00)	8.38(0.53,133.36)	2.47(0.54,11.27)	**5.60(1.24,25.31)**	5.40(0.89,32.91)	80
Complete Case	EM	0.38(0.06,2.30)	0.54(0.10,2.90)	7.13(0.45,113.79)	2.27(0.49,10.56)	**5.45(1.20,24.85)**	5.39(0.88,33.07)	80
					All Cases			
Naive	X_1	0.54(0.20,1.44)	2.11(0.78,5.76)	3.33(0.65,17.16)	2.42(0.88,6.72)	**3.20(1.17,8.73)**	**3.05(1.03,9.03)**	235
	X_2	**0.18(0.04,0.75)**	1.47(0.41,5.29)	5.33(0.71,39.82)	2.26(0.63,8.16)	**4.53(1.26,16.29)**	**7.33(1.43,37.66)**	160
	X_3	0.30(0.06,1.49)	0.47(0.08,2.58)	10.69(0.58,196.26)	2.52(0.54,11.77)	**5.54(1.19,25.72)**	5.16(0.84,31.66)	80
Full Data	EM	0.31(0.08,1.27)	1.94(0.69,5.39)	3.00(0.58,15.55)	2.38(0.85,6.68)	**3.18(1.15,8.78)**	2.70(0.89,8.16)	235

Notes: Here Z is ATM (high vs. low), W_1 is tumor size ($\geq$ 5cm vs. $<$ 5cm), W_2 is tumor grade (3 vs. 1/2), W_3 is LN status ($+$ vs. $-$), W_4 is LVI ($+$ vs. $-$) and W_5 is whether age at diagnosis is ≥ 65 years (yes vs. no).

Table 9: Estimated proportion of high ATM (p), sensitivity and specificity of the labs and associated 95% CIs using different methods in the breast cancer study.

			Evaluation of Labs					
			Sensitivity (95%CI)			Specificity (95%CI)		
Method	Detail	p(95%CI)	lab 1	lab 2	lab 3	lab 1	lab 2	lab 3
				Complete Cases				
Naive	X_1	0.42(0.32,0.54)	-	0.76(0.60,0.88)	0.79(0.63,0.90)	-	0.87(0.74,0.94)	0.80(0.66,0.89)
	X_2	0.40(0.30,0.51)	0.81(0.64,0.91)	-	0.84(0.68,0.93)	0.83(0.70,0.91)	-	0.81(0.68,0.90)
	X_3	0.45(0.34,0.56)	0.75(0.59,0.86)	0.75(0.59,0.86)	-	0.84(0.70,0.92)	0.89(0.75,0.95)	-
Ad Hoc	$Z^* = I(X. \geq 1)$	0.57(0.46,0.68)	0.74(-,-)	0.70(-,-)	0.78(-,-)	1.00(-,1.00)	1.00(-,1.00)	1.00(-,1.00)
	$Z^* = I(X. \geq 2)$	0.40(0.30,0.51)	0.91(0.75,0.97)	0.91(0.75,0.97)	0.94(0.78,0.98)	0.90(0.77,0.96)	0.94(0.82,0.98)	0.88(0.75,0.94)
	$Z^* = I(X. = 3)$	0.30(0.21,0.41)	1.00(0.00,-)	1.00(0.00,-)	1.00(0.00,-)	0.82(0.70,0.90)	0.86(0.74,0.93)	0.79(0.66,0.87)
Complete Case	EM	0.40(0.28,0.53)	0.88(0.65,0.97)	0.91(0.63,0.98)	0.91(0.64,0.98)	0.88(0.72,0.95)	0.94(0.76,0.99)	0.86(0.71,0.94)
				All Cases				
Naive	X_1	0.52(0.46,0.59)	-	0.84(0.75,0.91)	0.79(0.63,0.90)	-	0.82(0.72,0.89)	0.80(0.66,0.89)
	X_2	0.50(0.42,0.58)	0.81(0.71,0.88)	-	0.84(0.68,0.93)	0.85(0.75,0.91)	-	0.81(0.68,0.90)
	X_3	0.45(0.34,0.56)	0.75(0.59,0.86)	0.75(0.59,0.86)	-	0.84(0.70,0.92)	0.89(0.75,0.95)	-
Full Data	EM	0.49(0.39,0.58)	0.92(0.77,0.98)	0.95(0.80,0.99)	0.95(0.57,1.00)	0.85(0.73,0.93)	0.87(0.71,0.95)	0.85(0.70,0.93)

5 Discussion and Future Work

Here, we considered the issues in the analysis with misclassified covariates when there is no validation data and surrogate covariate data are incompletely observed. We proposed a likelihood-based approach to jointly modeling the response, surrogate covariates, and the latent covariate of interest. This likelihood is naturally optimized with an EM algorithm which yields estimates of the parameters of primary interest as well as the parameters for the operating characteristics of the diagnostic tests and the parameters indexing the probability model for the latent covariate. The proposed approach offers an efficient and straightforward approach to estimation and inference and the performance is relatively robust to the misclassification rates of the surrogate covariates and the sample size. The simulation studies demonstrated that the methods we proposed out-performance naive analyses based on the sub-sample of individuals with complete data on the surrogate methods since this results in a biased sub-sample even under the missing at random mechanism we have formulated for the surrogate covariates.

In the simulation studies, we have focused on the setting with a binary response and one binary covariate subject to misclassification. The proposed framework can be generalized naturally to accommodate different types of outcomes (e.g. censored or recurrent event responses) and multi-category latent variables if there are several diseases that may be represented in a sample. In more complex survival settings data may be truncated or there may be a cured fraction of individuals. Methods for handling such complications are also naturally cast into the framework of an EM algorithm and so the proposed methods can naturally be adapted to handle these complications. Finally, more elaborate regression models could also be formulated if there were interest in examining possible interaction effects between the disease status and auxiliary covariates.

The literature on methods for handling incomplete data has been burgeoning in recent years with much of the work involving methods which differ in their robustness and efficiency. The mean score method of Reilly and Pepe (1995) is a robust approach designed to weaken modeling assumptions about the covariate process typically through stratification on the available data including the response. This could be adopted in the present setting but challenges arise in defining strata when the data on the surrogate variables are incomplete. Use of inverse probability weighted complete case analyses (Rotnitzky et al., 1998) is often useful but in the present setting the latent variable is never known definitively; this likewise makes the use of augmented inverse probability weighted estimating functions (Bang and Robins, 2005) less natural. Multiple imputation (Schafer, 1999), however, could be adapted quite naturally for the current setting and it represents a convenient alternative to the likelihood-based procedure we describe. Future work would be worthwhile to explore the use of multiple imputation for the present problem.

Acknowledgements

The authors thank Dr Xiaolan Feng for permission to use the data from the study of patients with breast cancer at the Foothills Hospital in Calgary and Dr Haocheng Li for the introduction and stimulating the collaboration with Dr Feng.

Disclosure Statement

In accordance with Taylor & Francis policy and our ethical obligations as researchers, we are reporting that we have no financial and/or business interests and receive no funding from any company that may be affected by the research reported in the enclosed paper.

Funding

This work was supported by grants from the Natural Sciences and Engineering Research Council of Canada [grant numbers RGPIN 04594 and RGPIN 155849], the Breast Cancer Society of Canada, and startup funds at the University of Calgary.

Notes on Contributors

Hua Shen is an Assistant Professor of Biostatistics in the Department of Mathematics and Statistics and jointly appointed in the Natural Sciences Program at the University of Calgary. Her research interests are on the methodology development and statistical analysis of data arising from public health and medical research. Her current research focuses on developing statistical methods to analyze incomplete lifetime data involving latent processes arising in clinical trials and observational studies.

Richard J. Cook is Professor of Statistics in the Department of Statistics and Actuarial Science at the University ofWaterloo with cross-appointments at the School of Public Health and Health Systems at the University of Waterloo and the Faculty of Health Science at McMaster University. His research interests include the analysis of life history data, the design and analysis of clinical and epidemiological studies, and statistical methods for incomplete data.

Data Availability Statement

There is a data set associated with the paper. However, data sharing violates the valid privacy and security concerns of the principal investigators, therefore authors do not share or make open the data supporting the results or analyses presented in their paper.

References

Bang, H. and Robins, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973.

Buonaccorsi, J. (2010). *Measurement Error: Models, Methods, and Applications*. CRC Press.

Camp, R., Dolled-Filhart, M., and Rimm, D. (2004). X-tile: a new bio-informatics tool for biomarker assessment and outcome-based cut-point optimization. *Clinical Cancer Research*, 10(21):7252–7259.

Carroll, R., Ruppert, D., Crainiceanu, C., and Stefanski, L. (2006). *Measurement Error in Nonlinear Models: a Modern Perspective*. Chapman and Hall/CRC.

Chu, H., Cole, S. R., Wei, Y., and Ibrahim, J. G. (2009). Estimation and inference for case–control studies with multiple non–gold standard exposure assessments: with an occupational health application. *Biostatistics*, 10(4):591–602.

Dempster, A., Laird, N., and Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22.

Feng, X., Li, H., Dean, M., Wilson, H., Kornaga, E., Enwere, E., Tang, P., Paterson, A., Lees-Miller, S., Magliocco, A., et al. (2015). Low ATM protein expression in malignant tumor as well as cancer-associated stroma are independent prognostic factors in a retrospective study of early-stage hormone-negative breast cancer. *Breast Cancer Research*, 17(65):2817.

Feng, X., Li, H., Kornaga, E., Dean, M., Lees-Miller, S., Riabowol, K., Magliocco, A., Morris, D., Watson, P., Enwere, E., et al. (2016). Low Ki67/high ATM protein expression in malignant tumors predicts favorable prognosis in a retrospective study of early stage hormone receptor positive breast cancer. *Oncotarget*, 7(52):85798–85812.

Greenland, S. (1988). Variance estimation for epidemiologic effect estimates under misclassification. *Statistics in Medicine*, 7(7):745–757.

Grove, W., Andreasen, N., McDonald-Scott, P., Keller, M., and Shapiro, R. (1981). Reliability studies of psychiatric diagnosis: Theory and practice. *Archives of General Psychiatry*, 38(4):408–413.

Gustafson, P. (2003). *Measurement Error and Misclassification in Statistics and Epidemiology: Impacts and Bayesian Adjustments*. CRC Press.

Holcroft, C. and Spiegelman, D. (1999). Design of validation studies for estimating the odds ratio of exposure–disease relationships when exposure is misclassified. *Biometrics*, 55(4):1193–1201.

Liu, X. and Liang, K.-Y. (1991). Adjustment for non-differential misclassification error in the generalized linear model. *Statistics in Medicine*, 10(8):1197–1211.

Louis, T. (1982). Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 44(2):226–233.

Marshall, R. (1990). Validation study methods for estimating exposure proportions and odds ratios with misclassified data. *Journal of Clinical Epidemiology*, 43(9):941–947.

Meade, M., Guyatt, G., Cook, R., Groll, R., Kachura, J., Wigg, M., Cook, D., Slutsky, A., and Stewart, T. (2001). Agreement between alternative classifications of acute respiratory distress syndrome. *American Journal of Respiratory and Critical Care Medicine*, 163(2):490–493.

Morrissey, M. and Spiegelman, D. (1999). Matrix methods for estimating odds ratios with misclassified exposure data: extensions and comparisons. *Biometrics*, 55(2):338–344.

Perrot, S., Choy, E., Petersel, D., Ginovker, A., and Kramer, E. (2012). Survey of physician experiences and perceptions about the diagnosis and treatment of fibromyalgia. *BMC Health Services Research*, 12(356).

Reilly, M. and Pepe, M. S. (1995). A mean score method for missing and auxiliary covariate data in regression models. *Biometrika*, 82(2):299–314.

Rindskopf, D. and Rindskopf, W. (1986). The value of latent class analysis in medical diagnosis. *Statistics in Medicine*, 5(1):21–27.

Rotnitzky, A., Robins, J. M., and Scharfstein, D. O. (1998). Semiparametric regression for repeated outcomes with nonignorable nonresponse. *Journal of the American Statistical Association*, 93(444):1321–1339.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3):581–592.

Schafer, J. L. (1999). Multiple imputation: a primer. *Statistical Methods in Medical Research*, 8(1):3–15.

Shih, J. and Louis, T. (1995). Inferences on the association parameter in copula models for bivariate survival data. *Biometrics*, 51(4):1384–1399.

Spiegelman, D., Rosner, B., and Logan, R. (2000). Estimation and inference for logistic regression with covariate misclassification and measurement error in main study/validation study designs. *Journal of the American Statistical Association*, 95(449):51–61.

Walter, S. and Irwig, L. (1988). Estimation of test error rates, disease prevalence and relative risk from misclassified data: a review. *Journal of Clinical Epidemiology*, 41(9):923–937.

Yi, G. (2016). *Statistical Analysis with Measurement Error or Misclassification.* Springer.

Yi, G. and He, W. (2017). Analysis of case-control data with interacting misclassified covariates. *Journal of Statistical Distributions and Applications*, 4(16):1151.