Learning Models from Data with Measurement Error: Tackling Underreporting

Roy Adams[1] **Yuelong Ji**[2] **Xiaobin Wang**[2] **Suchi Saria**[1 3 4]

Abstract

Measurement error in observational datasets can lead to systematic bias in inferences based on these datasets. As studies based on observational data are increasingly used to inform decisions with real-world impact, it is critical that we develop a robust set of techniques for analyzing and adjusting for these biases. In this paper we present a method for estimating the distribution of an outcome given a binary exposure that is subject to underreporting. Our method is based on a missing data view of the measurement error problem, where the true exposure is treated as a latent variable that is marginalized out of a joint model. We prove three different conditions under which the outcome distribution can still be identified from data containing only error-prone observations of the exposure. We demonstrate this method on synthetic data and analyze its sensitivity to near violations of the identifiability conditions. Finally, we use this method to estimate the effects of maternal smoking and heroin use during pregnancy on childhood obesity, two import problems from public health. Using the proposed method, we estimate these effects using only subject-reported drug use data and refine the range of estimates generated by a sensitivity analysis-based approach. Further, the estimates produced by our method are consistent with existing literature on both the effects of maternal smoking and the rate at which subjects underreport smoking.

1. Introduction

Measurement error in observational datasets can lead to systematic bias in inferences based these datasets. As studies using observational data are increasingly used to inform decisions with real-world impact, it is critical that we develop a robust set of techniques for analyzing and adjusting for these biases. Bias can mean different things in different contexts, but here we use it in the statistical sense to refer to inferences that are wrong in a non-random way (Casella & Berger, 2002). Basing decisions on biased inferences can lead to real consequences and degrade trust in the use of observational data to inform decision making. For example, decisions about what resources should be allocated to preventing maternal drug use during pregnancy must be based on studies that rely on subject-reported drug use behaviors (Wang et al., 2002). Analyzing and accounting for various potential sources of bias remains an open problem, but one with important implications.

One reason that accounting for bias in observational data is a difficult problem is that it requires making further assumptions about what the sources and magnitudes of various biases might be. There is often substantial debate about the validity of these assumptions (e.g. Knape & Korner-Nievergelt (2015); Sólymos & Lele (2016); Knape & Korner-Nievergelt (2016)). As a result, it is frequently the case that observational studies relegate potential sources of bias to the discussion of limitations rather than performing quantitative bias analysis (Rothman et al., 2008). It is our view that, to the extent possible, quantitative bias analysis should be presented alongside quantitative measures of uncertainty when presenting results from observational studies. It is therefore critical that we develop a robust set of tools for analyzing and accounting for observational bias.

In this work we focus on bias caused by measurement error, the degree and direction of which depends heavily on the type of error (Carroll et al., 2006; Gustafson, 2003). There are three common approaches to dealing with measurement error (more details on these approaches are provided in Section 2). The first approach assumes that we have access to a data source containing error-free measurements (referred to as *validation data*). This data can be used to estimate the distribution of errors in order to adjust appropriately when

[1] Department of Computer Science, Johns Hopkins University [2] Center on the Life Origins of Disease, Department of Population, Family, and Reporductive Health, Johns Hopkins University Bloomberg School of Public Health [3] Department of Applied Math and Statistics, Johns Hopkins University [4] Bayesian Health. Correspondence to: Roy Adams <roy.james.adams@gmail.com>.

Proceedings of the 36th International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019. Copyright 2019 by the author(s).

error-free measurements are not available. When validation data is available, this approach is clearly preferable, but such data is often difficult or impossible to gather.

The second approach assumes that we can specify a small set of hypothetical error distributions which are then used to generate a corresponding range of inferences (Rothman et al., 2008). This set of hypothetical distributions can be specified using either domain knowledge or previous studies; however, if such knowledge is not available, then this type of sensitivity analysis may generate a wide range of inferences.

In the third approach, the error-free measurements are treated as unobserved variables, a model is specified that includes these variables, and inferences are made using only the observed error-prone data. If the modeling assumptions are correct, then this approach can give unbiased inferences without relying on validation data or specific knowledge about the amount of error present in the data.

In this work, we propose a method based on the third approach to account for a type of measurement error called *exposure misclassification*. Exposure misclassification occurs when we are interested in estimating the distribution of an outcome Y given a binary exposure A, but A is subject to measurement error[1]. For example, when estimating the effect of maternal drug use (exposure) on childhood obesity (outcome) using survey data, subjects have a tendency to *underreport* (source of error) whether or not they have used drugs.

The primary contribution of this work is a method for estimating the outcome distribution when we are only able to observe a version of the exposure that is subject to underreporting. Underreporting is common in problems involving survey-based observations of sensitive behaviors and, to our knowledge, this is the first method that allows unbiased estimation directly from such data. We prove three different assumptions under which such estimation is possible (Section 4), enabling flexible application of our method to different problem settings.

Using synthetic data, we show that this approach reduces estimation error compared with treating the error-prone observations as ground truth (Section 5). Finally, we use this approach to estimate the effects of maternal smoking and heroin use during pregnancy on childhood obesity using subject-reported drug use data. The effects of childhood obesity later in life can be severe and identifying potential causes of childhood obesity has significant public health implications. Due to underreporting error, obtaining unbiased estimates of these effects directly from survey data has not been possible and, to our knowledge, the contributions made in this article enabled the first reported estimate of the effect of maternal heroin use on childhood obesity. Estimates produced by the proposed method refine the range of estimates generated by a sensitivity analysis based approach and are consistent with existing literature on the effects of maternal smoking and the rate at which subjects underreport smoking (Section 6).

2. Background

Measurement error is encountered in a wide range of settings and there is extensive literature on various adjustment techniques. In this section, we contextualize our contributions with a review of the measurement error bias problem and some of the approaches to adjusting for this bias. We encourage the interested reader to seek out Carroll et al. (2006) or Gustafson (2003) for full treatments of this topic.

2.1. Measurement error bias

Suppose that we are interested in estimating the distribution of response Y given predictor A, denoted $p(Y|A)$. We may do this because we are interested in the parameters of this distribution or because we would like to predict Y from future values of A. However, suppose that rather than observing A in our training data, we observe $\tilde{A}$, a version of A that is corrupted by measurement error. In cases where $p(Y|A) \neq p(Y|\tilde{A})$, ignoring the measurement error in $\tilde{A}$ may lead us to a biased estimate of $p(Y|A)$. The specific effect using $\tilde{A}$ has on our estimate will depend on how $\tilde{A}$ relates to A and we will generally have to make some modeling assumptions about this relationship in order to adjust for it. For example, classical models for measurement error assume that $\tilde{A}$ equals A plus a noise variable that is independent of A (Carroll et al., 2006); however, these assumptions are not appropriate in many scenarios, including the one considered in this paper. Given a model for the error process, there are a number of ways we might go about adjusting for the bias caused by measurement errors.

2.2. Adjusting for measurement error bias

As discussed in the introduction, methods for adjusting for measurement error bias fall into three broad categories and choosing which method is appropriate for a particular problem will depend on what data and domain knowledge is available.

Auxiliary data: In the first category, it is assumed that some form of auxiliary data is available that allows us to estimate the error distribution directly. This auxiliary data may be validation data for which ground truth is available or it may be a second error-prone measurement of the same underlying variable. In the case of validation data, it is clear that the error distribution can be estimated; however, if the auxiliary data also contains errors, assumptions about the

[1] Measurement error in discrete variables is referred to as *misclassification*.

type and relationship of these errors are necessary to guarantee that the error distribution is estimable. For example, in Section 4.2 we prove for our model that the error distribution is estimable from two error-prone measurements which are independent given the true exposure. Once the error distribution is estimated, there are a number of ways one can correct for the missing ground truth observations such as imputation or sampling (Carroll et al., 2006). Using auxiliary data typically requires the least restrictive modeling assumptions and is therefore preferable when such data is available. It is unsurprising then that this approach forms the basis for much of the classic literature on measurement error (Carroll et al., 2006; Gustafson, 2003) and is used in modern approaches when possible (e.g. Pearl (2012)). In this work, however, we consider the case when no auxiliary data is available and these methods do not apply.

Sensitivity analysis: When we cannot directly estimate the error distribution from data, we can instead consider a range of hypothetical error distributions. If this range is small, then we can simply enumerate them, correcting for the missing ground truth observations as above, to generate a corresponding range of plausible inferences. This approach does not rely on any auxiliary data and can be applied to the cases we consider in this paper; however, if we do not have sufficient domain knowledge to specify a small set of error distributions, the resulting range of inferences may be quite large (as we demonstrate in Section 6).

Full likelihood: A third and less common approach is to specify a joint model for the target Y, the unobserved ground-truth predictor A, and the error-prone measurement $\tilde{A}$ and then marginalize A out of this joint model. This is often referred to as the *full likelihood* approach (Carroll et al., 2006; Thomas et al., 1993). Even if correctly specified, the parameters of this joint model cannot, in general, be estimated from the observed data without further assumptions about the structure of the joint distribution; however, there has been some work to identify such cases. For example, Rudemo et al. (1989) show that a particular class of non-linear medication dose response models can be estimated even if the true dose is subject to measurement error. Küchenhoff & Carroll (1997) show that a class of threshold regression models can be estimated from data with additive observation error in the predictors. The method presented in this paper is an example of the full likelihood approach applied the exposure misclassification problem.

3. Model

Suppose that we are interested in estimating the conditional distribution of a binary outcome $Y \in \{0, 1\}$ given a binary exposure binary exposure $A \in \{0, 1\}$ and a set of known confounding variables $X \in \mathbb{R}^{d2}$; however, rather than observing A, we observe an error-prone version $\tilde{A} \in \{0, 1\}$. For example, A might represent true drug use while $\tilde{A}$ represents subject-reported drug use. For reasonably high-dimensional covariates X, this is typically done using a parametric *outcome model* $p_\theta(Y|A, X)$ which, if we were able to observe A, could be estimated directly from data. Instead, our goal is to estimate $p_\theta(Y|A, X)$ from a dataset $\mathcal{D} = \{(x_i, \tilde{a}_i, y_y)\}_{i=1}^N$ consisting only of the outcome, covariates, and error-prone exposure observation.

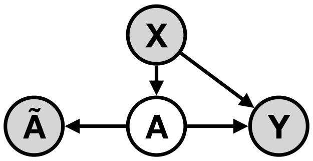

Figure 1. Proposed graphical model. Y is the outcome of interest, A is the true binary exposure, $\tilde{A}$ is the observed exposure, and X is a vector of covariates. Grey variables are observed and white variables are not.

In order to make this possible, we make two simplifying assumptions. First, we assume that the observed exposure $\tilde{A}$ is independent of the outcome and covariates given the true exposure or $\tilde{A} \perp X, Y|A$. In our drug use example, this means that the decision to misreport drug use is independent of known confounders such as age and income given the ground truth. When the conditional independence with Y holds, the measurement error is referred to as *non-differential*. This assumption is illustrated in the graphical model shown in Figure 1.

Second, we assume strict underreporting of the true exposure or $p(\tilde{A} = 1|A = 0) = 0$. For example, this assumption might encode our belief the non-users will not falsely report that they use drugs. Given these assumptions, we can write the conditional probability of Y and $\tilde{A}$ given X as[3]:

$$p(y, \tilde{a}|x) = \sum_a p(\tilde{a}|a)p(a|x)p(y|a, x)$$

where τ represents the underreporting rate. Finally, we will assume that the modeler has selected a parametric *propensity model* $p_\phi(a|x) \in \mathcal{P}_\Phi$ and a parametric *outcome model* $p_\theta(y|a, x) \in \mathcal{P}_\Theta$ for $p(a|x)$ and $p(y|a, x)$ respectively. We

[2] Under the assumptions of consistency, positivity, and conditional exchangeability, this distribution can be used to estimate the causal effect of A on Y (Hernán & Robins, 2018). In Sections 5 and 6, we will use it to estimate the risk difference or average treatment effect.

[3] When it does not cause confusion, we simplify $p(\tilde{A} = \tilde{a}|A = a)$ to $p(\tilde{a}|a)$.

parameterize the *error model* $p_\tau(\tilde{a}|a)$ as:

$$p_\tau(\tilde{a}=0|a=1) = 1 - p_\tau(\tilde{a}=1|a=1) = \tau$$
$$p_\tau(\tilde{a}=0|a=0) = 1 - p_\tau(\tilde{a}=0|a=0) = 1$$

We can estimate these parameters jointly by maximizing the following log conditional likelihood:

$$\mathcal{L}(\tau,\phi,\theta) = \sum_i \log \sum_a p_\tau(\tilde{a}_i|a) p_\phi(a|x_i) p_\theta(y_i|a,x_i) \tag{1}$$

Even in the case where ϕ and θ could be estimated from data containing the true exposure, we may not be able to estimate τ, ϕ, and θ from the available data without further assumptions. In the next section, we prove several such conditions under which the parameters of this model are estimable.

4. Identifiability

We prove three conditions under which the joint model $p_{\tau,\phi,\theta}(y,\tilde{a}|x)$ is identifiable[4]. These results allow us to use the model from Section 3 for each of the three bias analysis and adjustment approaches described in Section 2: sensitivity analysis, auxiliary data, and full likelihood.

4.1. Known τ

The validation data and sensitivity analysis approaches are based on the idea that, when $p_\tau(\tilde{a}|a)$ is known, $p_\theta(y|a,x)$ is estimable from data containing measurement error. We prove that this is true for the model presented in Section 3 allowing us to use this model to perform sensitivity analysis by fixing τ at various values and estimating ϕ and θ from the observed data. Specifically, we have the following result:

Proposition 1. *If τ is known and $p_{\phi,\theta}(y,a|x)$ is identifiable, then $p_{\tau,\phi,\theta}(y,\tilde{a}|x)$ is identifiable.*

Proof. By the identifiability of $p_{\phi,\theta}(y,a|x) = p_\theta(y|a,x)p_\phi(a|x)$, $(\phi,\theta) = (\phi',\theta')$ if and only if $p_{\phi,\theta}(y,a|x) = p_{\phi',\theta'}(y,a|x)$. Then for any x, y, and $\tau < 1$, marginalizing over A to map from $p_{\phi,\theta}(y,a|x)$ to $p_{\phi,\theta}(y,\tilde{a}|x)$ is an invertible linear map, so $p_\theta(y,\tilde{a}|x) = p_{\theta'}(y,\tilde{a}|x)$ if and only if $p_\theta(y,a|x) = p_{\theta'}(y,a|x)$. □

Unfortunately, it is frequently the case that we do not know τ and cannot specify a reasonably tight range of possible values, so we must estimate it from data.

4.2. Multiple error-prone observations

Next, we consider the case where τ is unknown, but we have access to a second error-prone exposure observation. For example, we may have access to both subject-reported drug use and lab measurements of biomarkers for drug use. We assume the second observation is also subject to the conditional independence and underreporting assumptions described in Section 3. In this scenario, both $\tilde{A}$ and τ are two dimensional vectors. In general, we cannot estimate τ, ϕ, and θ from the available data without further assumptions about $\tilde{A}$. One such assumption is that the two observations, $\tilde{A}_1$ and $\tilde{A}_2$, are conditionally independent given the true exposure A. When such data is available, we can use this result to estimate τ, ϕ, and θ with out any knowledge of the value of τ.

Proposition 2. *If $\tilde{A}_1 \perp \tilde{A}_2|A$ and $p_{\phi,\theta}(y,a|x)$ is identifiable, then $p_{\tau,\phi,\theta}(y,\tilde{a}|x)$ is identifiable.*

Proof. Assume for contradiction that the above condition holds and there exists $(\tau,\phi,\theta) \neq (\tau',\phi',\theta')$ such that $p_{\tau,\phi,\theta}(y,\tilde{a}|x) = p_{\tau',\phi',\theta'}(y,\tilde{a}|x)$ (i.e. $p_{\tau,\phi,\theta}(y,\tilde{a}|x)$ is not identifiable). Then the following equalities must hold:

$$(1-\tau_1)(1-\tau_2)\pi_\phi(x) = (1-\tau_1')(1-\tau_2')\pi_{\phi'}(x)$$
$$\tau_1(1-\tau_2)\pi_\phi(x) = \tau_1'(1-\tau_2')\pi_{\phi'}(x)$$
$$\tau_2(1-\tau_1)\pi_\phi(x) = \tau_2'(1-\tau_1')\pi_{\phi'}(x)$$

where $\pi_\phi(x) = p_\phi(A=1|x)$. By applying some algebra to these equalities, we have that $\tau_1 = \tau_1'$ and $\tau_2 = \tau_2'$. Finally, by Proposition 1, we have that $(\phi,\theta) = (\phi',\theta')$, which is a contradiction. □

Intuitively, we know there are no false positives, so the independence assumption $\tilde{A}_1 \perp \tilde{A}_2|A$ allows us to estimate τ_1 as the proportion of samples where $\tilde{A}_1 = 0$ among the samples where $\tilde{A}_2 = 1$ and vice versa.

4.3. Single error-prone observation

Finally, we consider the most difficult case: when we only have access to a single error-prone observation and no knowledge of τ. In this case, we can only guarantee identifiability of $p_{\tau,\phi,\theta}(y,\tilde{a}|x)$ by making further assumptions about the structure of the model. In Theorem 1, we prove one such condition which constrains the structure of the propensity model $p_\phi(a|x)$. Further, we prove in Corollary 1 that three of the most common Bernoulli regression models meet this condition, allowing us to use these models to estimate τ, ϕ, and θ from data containing only a single error-prone observation.

Theorem 1. *If $p_{\phi,\theta}(y,a|x)$ is identifiable and for all $\alpha \in [0,1)$ and $\phi,\phi' \in \Phi$, there exists x such that $p_\phi(A=1|x) \neq \alpha p_{\phi'}(A=1|x)$, then $p_{\tau,\phi,\theta}(y,\tilde{a}|x)$ is identifiable.*

[4] A parametric model $p_w \in \mathcal{P}_\mathcal{W}$ is *identifiable* if $p_w = p_{w'} \implies w = w'$ for all $w, w' \in \mathcal{W}$. This is equivalent to saying that, in the limit of infinite data from p_w, we could identify the parameters that gave rise to this data. Importantly, this is a property of a model and is independent of the data.

Proof. Assume for contradiction that the above condition holds and there exists $(\tau, \phi, \theta) \neq (\tau', \phi', \theta')$ such that $p_{\tau,\phi,\theta}(y, \tilde{a}|x) = p_{\tau',\phi',\theta'}(y, \tilde{a}|x)$ (i.e. $p_{\tau,\phi,\theta}(y, \tilde{a}|x)$ is not identifiable). By assumption, we have

$$\begin{aligned}&(1-\tau)p_\phi(A=1|x)p_\theta(y|A=1,x)\\&=(1-\tau')p_{\phi'}(A=1|x)p_{\theta'}(y|A=1,x).\end{aligned}$$

It must also be the case that $p_\theta(y|A = 1, x) = p_{\theta'}(y|A = 1, x)$ since otherwise at least one of $p_\theta(y|A = 1, x)$ or $p_{\theta'}(y|A = 1, x) = \frac{(1-\tau)p_\phi(A=1|x)}{(1-\tau')p_{\phi'}(A=1|x)}p_\theta(y|A = 1, x)$ would not sum to one. Therefore, $p_\theta(y|A = 1, x) = p_{\theta'}(y|A = 1, x)$ and

$$(1-\tau)p_\phi(A=1|x) = (1-\tau')p_{\phi'}(A=1|x)$$

Without loss of generality, assume that $\tau < \tau'$. Then, $p_\phi(A = 1|x) = \alpha p_\phi(A = 1|x)$ where $\alpha = \frac{1-\tau'}{1-\tau} \in [0, 1)$ which is a contradiction. □

This result is somewhat technical, but intuitively, it says that if there exists ϕ, ϕ', and $\alpha \in [0, 1)$ such that $p_\phi(A = 1|x) = \alpha p_{\phi'}(A = 1|x)$ then it will be impossible to distinguish $(\alpha\tau, \phi)$ from (τ, ϕ') given the observed data. Fortunately, this condition holds for three common Bernoulli regression models.

Corollary 1. *If $A \not\perp X$, then a Bernoulli regression model $p_\phi(A = 1|x) = \Psi^{-1}(\phi x)$ with a logit, probit, or complementary log-log (cloglog) link function Ψ satisfies the identifiability condition in Theorem 1.*

Proof. A Bernoulli regression model of the form $p_\phi(A = 1|x) = \Psi^{-1}(\phi x)$ violates the condition in Theorem 1 if and only if there exists $\alpha \in [0, 1)$, ϕ, and ϕ' such that

$$\Psi(\alpha\Psi^{-1}(\phi x)) = \phi' x$$

For logistic regression, this is true if

$$\log(\alpha) - \log(1 - \alpha + e^{-\phi x}) = \phi' x$$

For all $\alpha < 1$, the function $\log(1 - \alpha + e^{-x})$ is non-linear in x, so this equality can only be true if ϕx and $\phi' x$ are constants which is true only when $A \perp X$. The same argument can be applied to the probit and complementary log-log link functions (see the supplementary materials for details). □

These three results allow us to estimate the outcome model directly from the observed data in a variety of scenarios. Which result is most applicable will depend on the particular modeling problem. For example, if τ is known based on previous literature, then we should use Proposition 1 as this result makes the fewest assumptions about the structure of the model. Equipped with these results, we apply the full likelihood approach to estimation problems using both synthetic and real data.

5. Synthetic experiments

We conducted a series of synthetic experiments to demonstrate the behavior of the full likelihood approach based on Theorem 1 under varied data conditions and near violations of the identifiability conditions. Unless otherwise stated, synthetic data was generated using the following model:

$$\begin{aligned}X_i &\sim \mathcal{N}(\mathbf{0}, \mathbf{I})\\A_i|X_i &\sim \text{Bern}(\text{expit}(\phi_0 + \phi X_i))\\Z_i &\sim \text{Bern}(1-\tau)\\\tilde{A}_i &= Z_i A_i\\Y_i|A_i, X_i &\sim \text{Bern}(\text{expit}(\theta_0 + \theta_X X_i + \theta_A A_i))\end{aligned} \tag{2}$$

where ϕ and θ were both sampled from a standard normal and θ_A was set to 1.0. For any $\theta_X \neq \mathbf{0}$, this data generating process satisfies the identifiability condition of Corollary 1, so the parameters should be identifiable using a single error-prone observation of the exposure. We evaluated a version of the full likelihood method based on Theorem 1 where both $p_\phi(A|X)$ and $p_\theta(Y|A, X)$ were logistic regression models and we maximized the conditional likelihood in Equation 1 using L-BFGS (”adjusted”). We compared against a logistic regression model estimated under the assumption of no observation error, that is $\tilde{A} = A$ (”not adjusted”).

In all experiments the target estimand was the risk difference (RD) which we estimated as[5]:

$$\begin{aligned}RD \approx \frac{1}{N}\sum_{i=1}^{N} &p_{\hat{\theta}}(Y=1|A=1, X=x_i)\\&- p_{\hat{\theta}}(Y=1|A=0, X=x_i)\end{aligned} \tag{3}$$

where $\hat{\theta}$ is the estimated value of θ.

While adjusting for measurement error may reduce estimator bias, it may also increase variance[6]. In this way, adjusting for noise in the observation process can be viewed a bias/variance trade-off. Accordingly, we compare the two approaches in terms of mean squared error (MSE) which allows us to judge whether adjusting for measurement errors makes a favorable bias/variance trade-off. Our expectation is that adjusting for measurement error bias leads to reduced estimation error in all cases where the modeling assumptions are satisfied. MSE was approximated by averaging the squared-error over 1,000 synthetic datasets.

In our first experiment (Figure 2 (a)), we evaluated both approaches while varying the true underreporting rate τ between 0 and 0.8 and keeping the dataset size fixed at $1,000$.

[5] Under the assumptions of consistency, positivity, and conditional exchangeability, the risk difference can be interpreted as a measure of the causal effect of A on Y (Hernán & Robins, 2018).

[6] Adjusting for measurement error in the observations reduces the effective sample size.

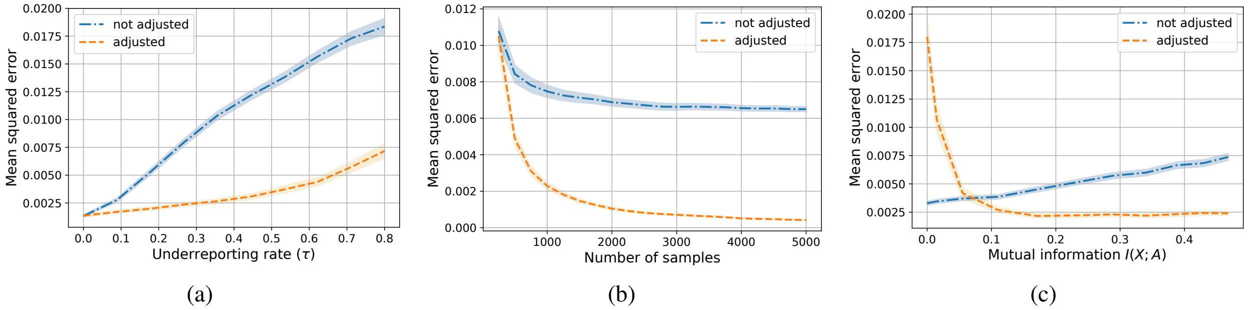

Figure 2. Experimental results on synthetic data. Data was generated from the model described in Equation 2 with specific parameters varied in each experiment. All plots compare the MSE of the proposed method to a model fit under the assumption that there is no measurement error. Error bars show 95% confidence intervals. In order, the plots show MSE as a function of (a) the true underreporting rate, (b) the sample size, and (c) the mutual information between A and X.

In all cases, adjusting for measurement error reduced MSE, with the gap in MSE increasing as the true underreporting rate increased. This reflects the intuitive idea that less measurement error results in less bias and therefore, there is less to be gained by adjusting for this bias.

In our second experiment (Figure 2 (b)), we varied the dataset size between 250 and $5,000$ while keeping τ fixed at 0.25. For small datasets, where we expect variance to swamp bias, both approaches resulted in comparable MSE; however, as the data size increased, the adjusted estimate converged to zero MSE while the unadjusted estimate converged to a fixed bias. This result demonstrates the importance of adjusting for bias when using large datasets so that we do not become overly confident in the wrong inference.

In our third experiment (Figure 2 (c)), we tested the sensitivity to the assumption in Corollary 1 by evaluating how the MSE of the adjusted model increases as the dependence between A and X decreases. Figure 2 (c) shows this dependence in terms of mutual information. In this experiment, we left τ fixed at 0.25 and the data size fixed at $1,000$ while varying the magnitude of ϕ_X. Specifically, we multiplied the weights ϕ_X by a scaler c ranging from 0 to 1 (ϕ_0 was left fixed so that $\mathbf{E}[A]$ remained constant). As we would expect based on Corollary 1, the MSE of the adjusted model increases significantly as the mutual information between X and A approaches zero, with the unadjusted model performing better for some mutual information values greater than zero. An important implication of this result is that if we see a large difference in the estimated variance for the full likelihood and unadjusted approaches, this may be an indication of a near violation of the identifiability conditions. In the next section, we apply the full likelihood method to an effect estimation problem from public health.

6. Boston birth cohort

Childhood overweight and obesity (COWO) affects over 30% of children in the United States (Ogden et al., 2014) and is associated with a variety of chronic adulthood diseases such as adult obesity (Suchindran et al., 2010) and stroke (Lawlor & Leon, 2005). Understanding the early childhood and pre-birth exposures that lead to COWO is a critical for developing interventions that may reduce COWO rates. One such potential exposure is maternal drug use during pregnancy which has been linked to several adverse birth outcomes (Wang et al., 2002; Robison et al., 2012; Whiteman et al., 2014).

In this section, we use the full likelihood method based on Theorem 1 to estimate the effects of smoking and heroin use on COWO. It is difficult to accurately measure drug use in large populations, so we are often forced to rely on subject-reported drug use, which is known to be unreliable (Patrick et al., 1994; Gorber et al., 2009; Boyd et al., 1998). Both the effect of smoking on COWO and the rate at which study subjects underreport smoking have previously been studied (e.g von Kries et al. (2007); Gorber et al. (2009)) which allows us check our estimates of these quantities against existing literature. The effect of heroin use on childhood obesity has, to our knowledge, never been studied and we include it as a demonstration of our method on an important unanswered question.

Data: To estimate these effects, we use data from the Boston Birth Cohort, a longitudinal dataset tracking health markers from mothers and children. For each mother/child pair i, the outcome Y_i is equal to one if the most recent measurement of the child's body mass index (BMI) is above the 85th percentile, and the true exposure A_i represents whether or not the mother used the substance in question during pregnancy. The complete dataset contains 8,507 mother/child pairs and we have child BMI measurements for 2,763 of those pairs. In the cases of both smoking and heroin use,

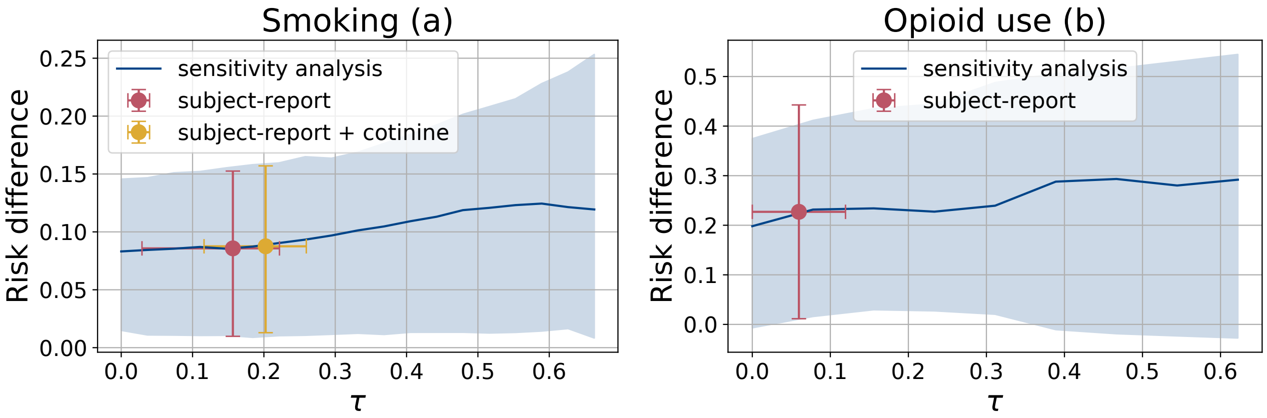

Figure 3. Risk difference estimates for (a) smoking and (b) heroin use. The blue lines show the range of risk difference estimates generated using a sensitivity analysis approach with the shaded region indicating 95% confidence intervals. Dots represent risk difference estimated generated using a full likelihood method with y-axis error bars indicating 95% confidence intervals for the risk difference estimate and x-axis error bars indicating 95% confidence intervals for the estimate of τ.

we have access to subject-reported substance use indicators which we use as error-prone exposure measurements $\tilde{A}_i$. In the case of smoking, we also have measurements of blood cotinine levels (a nicotine metabolite) at the time of delivery for 1,333 of the mothers. As a second error-prone indicator of maternal smoking, we used a binary variable equal to one if this cotinine measurement was available and above the 95th percentile. In both cases, we adjusted for the covariates X_i given in Wang et al. (2016), which include (among others) age, income, and other substance use indicators.

Model: In all cases, we used logistic regression models for both $p_\phi(a|x)$ and $p_\theta(y|a,x)$. Using a logistic regression model, we estimated the mutual information between X and $\tilde{A}$ at 0.17 for smoking and 0.14 for heroin use, which gives us some evidence that $A \not\perp X$, as required by Corollary 1. We estimated τ, ϕ, and θ by maximizing the log conditional likelihood in Equation 1 using L-BFGS. As in our synthetic experiments (Section 5) our target estimand was the risk difference, which we estimated according to Equation 3.

The effect of smoking on childhood obesity: When estimating the effect of smoking on childhood obesity, we compared three methods: a sensitivity analysis where τ was fixed at values ranging from 0 and 0.65, the full likelihood method using both subject-reported smoking and cotinine levels, and the full likelihood method using only subject-reported smoking. The resulting estimates along with 95% confidence intervals are shown in Figure 3 (a). The sensitivity analysis approach gave a range of estimated risk differences between 0.083 and 0.124 while the two full likelihood methods gave similar estimates of 0.086 and 0.088. Additionally, the two joint estimation approaches gave estimates for the underreporting rate of 0.157 and 0.202.

Three previous studies on the effect of maternal smoking on childhood obesity reported adjusted odds ratios of 1.60 (von Kries et al., 2007), 1.43 (Von Kries et al., 2002), and 1.58 (Toschke et al., 2002). The two full likelihood methods estimated the adjusted odds ratio at 1.45 and 1.47, fully in line with the previous literature. Further, a review of previous studies on underreporting rates for smoking found the majority of studies reported underreporting rates between 0.01 and 0.47 which is in line with our estimated underreporting rates of 0.157 and 0.202 (Gorber et al., 2009).

The effect of heroin use on childhood obesity: When estimating the effect of maternal heroin use on childhood obesity, we compared a sensitivity analysis approach to the full likelihood method using subject-reported heroin use as an error-prone exposure measurement. The resulting estimates are shown in Figure 3 (b). The sensitivity analysis approach gave a range of estimated risk differences between 0.198 and 0.293 while the joint estimation approach estimated the risk difference at 0.227 and the underreporting rate at 0.060. Using the full likelihood method gives us evidence that the underreporting rate for heroin use is relatively low and the effect of maternal heroin use on COWO is on the lower end of the range given by the sensitivity analysis approach. Additionally, at the upper end of the sensitivity analysis range, the 95% confidence intervals grow to include zero. The full likelihood method, on the other hand, gives a confidence interval that excludes zero, giving us stronger evidence that an effect does, in fact, exist.

7. Related work

In this section, we highlight two relevant lines of work. In machine learning, the full likelihood approach has been used to learn classifiers from data in which the target label is subject to measurement error. One of the common approaches to solving this problem is to treat the true label as a latent variable which is marginalized out of the model (Yan et al., 2010; Raykar et al., 2009; Jin & Ghahramani, 2002; Mnih & Hinton, 2012). Unlike the exposure misclassification problem, the goal in the label noise problem is generally prediction, so models are evaluated empirically in terms of prediction accuracy with no guarantees about model identifiability.

The full likelihood approach has also been applied to a problem in computational ecology called *occupancy modeling*. In the occupancy modeling problem the goal is to estimate the probability that a region is inhabited by a particular species given features of the region. Occupancy models are typically estimated from survey data which often shares the same strict underreporting property that we discussed in Section 3. That is, if a species is observed, we assume the site is, in fact, occupied by the species. This problem differs from ours in that the measurement error is present in the target Y rather than the exposure A, but the marginal likelihood approach used Sólymos et al. (2012) is very similar to ours. In Sólymos et al. (2012) and the resulting discussion (Knape & Korner-Nievergelt, 2015; Sólymos & Lele, 2016; Knape & Korner-Nievergelt, 2016) the authors present identifiability conditions for their model and the results presented in Section 4 are extensions of these conditions to the problem of exposure misclassification.

8. Discussion

Measurement error is widespread in observational data and can lead to inferential bias with real-world implications. We argued that performing quantitative bias analysis and adjustment should be a first order concern for researchers using observational data and we presented a new method to adjust for the bias caused by exposure underreporting. We showed that this method can be used in a variety of scenarios, including when only a single error-prone exposure observation is available, a capability that did not exist before.

We demonstrated on synthetic data that this method can potentially reduce estimation error by a significant amount relative to ignoring measurement error, but may be sensitive to near violations of the modeling assumptions. Finally, we used this method to estimate the effects of maternal smoking and opioid use during pregnancy on childhood obesity. Our method refined the range of estimates given by a sensitivity analysis approach and, in the case of smoking, resulted in estimates matching previous literature.

There are a three additional points about this method that we would like to highlight. First, we do not view our method as a replacement to traditional sensitivity analysis, but rather as a complementary approach that can be used to improve and refine the range of estimates generated by sensitivity analysis. Specifically, when performing bias analysis, we recommend generating the types of plots shown in Figure 3 which combine results from both full likelihood and sensitivity analysis approaches. By analyzing the data under multiple sets of assumptions, we make our results more robust to violations of any single assumption[7].

Second, while Theorem 1 allows us to use several common Bernoulli regression models it is worth considering which models are excluded by this condition. An example of one such model is a scaled logistic regression model of the form $p_\phi(A = 1|x) = \alpha \text{expit}(\phi x)$ where $\alpha \in [0, 1]$. In this model, $p_\phi(A = 1|x)$ saturates at a value $\alpha \leq 1$ which allows us to control the maximum value of $p_\phi(A = 1|x)$ independently from sharpness of the decision boundary. As suggested by Sólymos & Lele (2016), such a model may be approximated by using logistic regression paired with a polynomial basis expansion of the covariates. While this type of feature expansion does not violate the condition in Corollary 1, it may increase estimator variance to the point where regularization becomes necessary to stabilize estimation. For a discussion this stabilization and further models that violate this condition, see Knape & Korner-Nievergelt (2015), Sólymos & Lele (2016), and Knape & Korner-Nievergelt (2016).

Finally, the condition set forth in Corollary 1 that $A \not\perp X$ may seem somewhat counterintuitive in a causal inference setting where randomized trials in which $A \perp X$ are often considered the gold standard. While high mutual information between between X and A is beneficial for recovery of A from error prone observations, it can also lead to near positivity violations. We recommend that practitioners visualize the distribution of $S = p(A = 1|X)$ to check for distributions concentrated around zero and one as is often done when using propensity score methods.

There are several potential future directions for this work. First, relaxing the conditional independence and strict underreporting assumptions from Section 3 would allow broader application of this method. For example, Sólymos et al. (2012) suggest an alternative to Corollary 1 for the occupancy modeling problem that can be used when $\tilde{A} \not\perp X|A$. Similarly, while we think it unlikely that we can eliminate all assumptions about the error distribution, generalizing the strict underreporting assumption to allow for false positives is necessary for many potential applications. Second, Rothman et al. (2008) suggest using semi-bayesian methods instead of sensitivity analysis to generate a distribution over inferences. In a similar vein, a fully Bayesian version of the approach presented in this paper would allow the modeler to refine the sensitivity analysis approach in a principled way while potentially stabilizing the full likelihood approach in the case of near violations of the identifiability conditions.

[7] This is the very principle underlying sensitivity analysis.

Acknowledgements

We thank the anonymous reviewers for their helpful feedback. This work was supported by funding from DARPA YFA number D17AP00014 and NSF SCH number 1418590. The Boston Birth Cohort (the parent study) is supported in part by the National Institutes of Health (NIH) under grants number R01HD086013 and 2R01HD041702; and by the Health Resources and Services Administration (HRSA) of the U.S. Department of Health and Human Services (HHS) under grant number R40MC27443 and cooperative agreement UJ2MC31074. This information or content and conclusions are those of the authors and should not be construed as the official position or policy of, nor should any endorsements be inferred by NIH, HRSA, HHS, NSF, DARPA, or the U.S. Government. Dr. Saria is founder and the University and her both hold equity in Bayesian Health. She is or has recently served as an advisor to the following biotechnology companies: Patient Ping, Novartis, Scripps Translational Research. This arrangement has been reviewed and approved by the Johns Hopkins University in accordance with its conflict of interest policies.

References

Boyd, N. R., Windsor, R. A., Perkins, L. L., and Lowe, J. B. Quality of measurement of smoking status by self-report and saliva cotinine among pregnant women. *Maternal and child health journal*, 2(2):77–83, 1998.

Carroll, R. J., Ruppert, D., Crainiceanu, C. M., and Stefanski, L. A. *Measurement error in nonlinear models: a modern perspective*. Chapman and Hall/CRC, 2006.

Casella, G. and Berger, R. L. *Statistical inference*, volume 2. Duxbury Pacific Grove, CA, 2002.

Gorber, S. C., Schofield-Hurwitz, S., Hardt, J., Levasseur, G., and Tremblay, M. The accuracy of self-reported smoking: a systematic review of the relationship between self-reported and cotinine-assessed smoking status. *Nicotine & tobacco research*, 11(1):12–24, 2009.

Gustafson, P. *Measurement error and misclassification in statistics and epidemiology: impacts and Bayesian adjustments*. CRC Press, 2003.

Hernán, M. A. and Robins, J. M. *Causal inference*. 2018.

Jin, R. and Ghahramani, Z. Learning with multiple labels. In *Advances in neural information processing systems*, pp. 897–904, 2002.

Knape, J. and Korner-Nievergelt, F. Estimates from nonreplicated population surveys rely on critical assumptions. *Methods in Ecology and Evolution*, 6(3):298–306, 2015.

Knape, J. and Korner-Nievergelt, F. On assumptions behind estimates of abundance from counts at multiple sites. *Methods in Ecology and Evolution*, 7(2):206–209, 2016.

Küchenhoff, H. and Carroll, R. Segmented regression with errors in predictors: semi-parametric and parametric methods. *Statistics in Medicine*, 16(2):169–188, 1997.

Lawlor, D. A. and Leon, D. A. Association of body mass index and obesity measured in early childhood with risk of coronary heart disease and stroke in middle age: findings from the aberdeen children of the 1950s prospective cohort study. *Circulation*, 111(15):1891–1896, 2005.

Mnih, V. and Hinton, G. E. Learning to label aerial images from noisy data. In *Proceedings of the 29th International conference on machine learning (ICML-12)*, pp. 567–574, 2012.

Ogden, C. L., Carroll, M. D., Kit, B. K., and Flegal, K. M. Prevalence of childhood and adult obesity in the united states, 2011-2012. *Jama*, 311(8):806–814, 2014.

Patrick, D. L., Cheadle, A., Thompson, D. C., Diehr, P., Koepsell, T., and Kinne, S. The validity of self-reported smoking: a review and meta-analysis. *American journal of public health*, 84(7):1086–1093, 1994.

Pearl, J. On measurement bias in causal inference. *arXiv preprint arXiv:1203.3504*, 2012.

Raykar, V. C., Yu, S., Zhao, L. H., Jerebko, A., Florin, C., Valadez, G. H., Bogoni, L., and Moy, L. Supervised learning from multiple experts: whom to trust when everyone lies a bit. In *Proceedings of the 26th Annual international conference on machine learning*, pp. 889–896. ACM, 2009.

Robison, R. G., Kumar, R., Arguelles, L. M., Hong, X., Wang, G., Apollon, S., Bonzagni, A., Ortiz, K., Pearson, C., Pongracic, J. A., et al. Maternal smoking during pregnancy, prematurity and recurrent wheezing in early childhood. *Pediatric pulmonology*, 47(7):666–673, 2012.

Rothman, K. J., Greenland, S., Lash, T. L., et al. Modern epidemiology. 2008.

Rudemo, M., Ruppert, D., and Streibig, J. Random-effect models in nonlinear regression with applications to bioassay. *Biometrics*, pp. 349–362, 1989.

Sólymos, P. and Lele, S. R. Revisiting resource selection probability functions and single-visit methods: Clarification and extensions. *Methods in Ecology and Evolution*, 7(2):196–205, 2016.

Sólymos, P., Lele, S., and Bayne, E. Conditional likelihood approach for analyzing single visit abundance survey data in the presence of zero inflation and detection error. *Environmetrics*, 23(2):197–205, 2012.

Suchindran, C., North, K. E., Popkin, B. M., Gordon-Larsen, P., et al. Association of adolescent obesity with risk of severe obesity in adulthood. *Jama*, 304(18):2042–2047, 2010.

Thomas, D., Stram, D., and Dwyer, J. Exposure measurement error: influence on exposure-disease relationships and methods of correction. *Annual review of public health*, 14(1):69–93, 1993.

Toschke, A., Koletzko, B., Slikker, W., Hermann, M., and von Kries, R. Childhood obesity is associated with maternal smoking in pregnancy. *European journal of pediatrics*, 161(8):445–448, 2002.

Von Kries, R., Toschke, A. M., Koletzko, B., and Slikker Jr, W. Maternal smoking during pregnancy and childhood obesity. *American journal of epidemiology*, 156(10):954–961, 2002.

von Kries, R., Bolte, G., Baghi, L., Toschke, A. M., and Group, G. S. Parental smoking and childhood obesity—is maternal smoking in pregnancy the critical exposure? *International journal of epidemiology*, 37(1):210–216, 2007.

Wang, G., Johnson, S., Gong, Y., Polk, S., Divall, S., Radovick, S., Moon, M., Paige, D., Hong, X., Caruso, D., et al. Weight gain in infancy and overweight or obesity in childhood across the gestational spectrum: a prospective birth cohort study. *Scientific reports*, 6:29867, 2016.

Wang, X., Zuckerman, B., Pearson, C., Kaufman, G., Chen, C., Wang, G., Niu, T., Wise, P. H., Bauchner, H., and Xu, X. Maternal cigarette smoking, metabolic gene polymorphism, and infant birth weight. *Jama*, 287(2):195–202, 2002.

Whiteman, V. E., Salemi, J. L., Mogos, M. F., Cain, M. A., Aliyu, M. H., and Salihu, H. M. Maternal opioid drug use during pregnancy and its impact on perinatal morbidity, mortality, and the costs of medical care in the united states. *Journal of pregnancy*, 2014, 2014.

Yan, Y., Rosales, R., Fung, G., Schmidt, M. W., Valadez, G. H., Bogoni, L., Moy, L., and Dy, J. G. Modeling annotator expertise: Learning when everybody knows a bit of something. In *AISTATS*, pp. 932–939, 2010.