Published in final edited form as:

Health Serv Outcomes Res Methodol. 2017 June ; 17(2): 101–112. doi:10.1007/s10742-016-0149-5.

Accounting for misclassification in electronic health records-derived exposures using generalized linear finite mixture models

Rebecca A. Hubbard, Eric Johnson, Jessica Chubak, Karen J. Wernli, Aruna Kamineni, Andy Bogart, and Carolyn M. Rutter

Department of Biostatistics & Epidemiology, University of Pennsylvania, Philadelphia, Pennsylvania (Rebecca A. Hubbard); Group Health Research Institute, Seattle, Washington (Eric Johnson, Jessica Chubak, Karen J. Wernli, Aruna Kamineni); Department of Epidemiology, University of Washington, Seattle, Washington (Jessica Chubak); RAND Corporation, Santa Monica, California (Andy Bogart, Carolyn M. Rutter)

Abstract

Exposures derived from electronic health records (EHR) may be misclassified, leading to biased estimates of their association with outcomes of interest. An example of this problem arises in the context of cancer screening where test indication, the purpose for which a test was performed, is often unavailable. This poses a challenge to understanding the effectiveness of screening tests because estimates of screening test effectiveness are biased if some diagnostic tests are misclassified as screening. Prediction models have been developed for a variety of exposure variables that can be derived from EHR, but no previous research has investigated appropriate methods for obtaining unbiased association estimates using these predicted probabilities. The full likelihood incorporating information on both the predicted probability of exposure-class membership and the association between the exposure and outcome of interest can be expressed using a finite mixture model. When the regression model of interest is a generalized linear model (GLM), the expectation-maximization algorithm can be used to estimate the parameters using standard software for GLMs. Using simulation studies, we compared the bias and efficiency of this mixture model approach to alternative approaches including multiple imputation and dichotomization of the predicted probabilities to create a proxy for the missing predictor. The mixture model was the only approach that was unbiased across all scenarios investigated. Finally, we explored the performance of these alternatives in a study of colorectal cancer screening with colonoscopy. These findings have broad applicability in studies using EHR data where gold-standard exposures are unavailable and prediction models have been developed for estimating proxies.

Corresponding author: Rebecca A. Hubbard, 604 Blockley Hall, 423 Guardian Dr, Philadelphia, PA 19104, phone: 215-898-311, fax: 215-473-5865, rhubb@upenn.edu.

Compliance with Ethical Standards:

Ethical approval: All procedures performed in studies involving human participants were in accordance with the ethical standards of the Group Health Institutional Review Board and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards. This article does not contain any studies with animals performed by any of the authors.

Informed consent: The Group Health Institutional Review Board approved a waiver of consent for this study.

Conflict of Interest: All authors declare that they have no conflicts of interest.

Keywords

Colorectal cancer; Electronic health records; Mixture model; Multiple imputation; Screening

1 Introduction

Data from electronic health records (EHR) have the potential to substantially accelerate the pace of health services research while reaching patient sub-groups often excluded from traditional, designed research studies. However, EHR data are often missing gold-standard information on key variables. Much effort has focused on the development of prediction models for such imperfectly ascertained variables (Richesson, Smerek 2015), with relatively little effort focusing on appropriate statistical methods for incorporating the resultant predictions into subsequent analyses.

Evaluation of the effectiveness of cancer screening tests provides an example of a research area where EHR data can be particularly informative. EHR data provide access to information on a large population, facilitating study of rare outcomes such as cancer, and allow for evaluation of screening test effectiveness in community practice. However, EHR-based cancer screening research is hampered by lack of availability of a key exposure variable, the purpose for conducting the test, referred to as the test indication.

Missing or erroneous information about the test indication poses a critical challenge to the evaluation of screening test outcomes using observational data (Weiss 1998; Weiss et al. 1992) and is common in research using EHR data. A test is defined as having a “screening indication” only if it is performed for the purposes of disease detection in the absence of signs or symptoms of disease. For instance, in the case of colorectal cancer, colonoscopy is recommended as a screening test for individuals over the age of 50 once per decade (US Preventive Services Task Force 2008). Colonoscopy is also used as a diagnostic test in individuals with signs or symptoms of colorectal cancer. Analyzing the association between test indication and cancer characteristics or outcomes thus represents an example of an application where EHR-based studies are missing gold-standard information on the exposure of interest.

Prediction models have been developed for many EHR-derived characteristics, including test indication (El-Serag et al. 2006; Fisher et al. 2011; Adams et al. 2015; Richesson, Smerek 2015). One approach to estimating the effect of these imperfectly ascertained characteristics on outcomes of interest is to directly use the prediction model-based probability of the characteristic as the exposure variable in a regression model. Previous research has demonstrated that using predicted probabilities in place of an unknown categorical exposure in linear regression models results in unbiased estimates with little loss of efficiency (McCaffrey, Elliott 2008; Elliott et al. 2009). However, implications for using predicted probabilities in non-linear regression have not been studied.

Mixture models provide a convenient representation of the likelihood when data include a missing categorical covariate. The likelihood can be formulated as a mixture distribution, with mixing probabilities, the probabilities assigned to each component of the mixture

likelihood, based on predicted probabilities of class membership. Likelihood-based estimation via the expectation-maximization (EM) algorithm can be used when the missing variable is an exposure variable in a generalized linear model (GLM) (Jansen 1993), and covariates can be used to improve prediction of mixing probabilities (Thompson et al. 1998). Mixture models have been broadly used to address the case where members of a population are known to come from distinct sub-populations but specific class membership is unknown (Redner, Walker 1984; Hunt, Jorgensen 2003). The goal of the current paper is to compare the performance of the mixture model approach to naïve approaches currently used in the analysis of imperfect exposures obtained from EHR data, and to demonstrate the relative bias and precision of these alternatives. We focus on the special case where probabilities of class membership are available from a prior validation study and the goal of estimation is to obtain parameter estimates from a GLM. This is a commonly encountered situation in studies of EHR data where prior validation studies provide models for predicted probabilities of class membership and GLMs, such as the logistic regression model, are used to estimate the relationship between class membership and outcomes.

We use the case of imperfectly ascertained screening test indication as a motivating example for comparison of alternative methods. We present four alternative statistical approaches and compare their performance using simulation methods. We then demonstrate the application of these methods in an empirical study of stage at diagnosis among colorectal cancers detected by screening vs. diagnostic colonoscopy using data from Group Health, an integrated healthcare system in Washington State. Because cancer screening tests work by identifying disease at an earlier and more treatable stage, determining if late stage cancers are rarer in the population of screen-detected cancers compared to diagnostically detected cancers, referred to as “stage shift”, provides a measure of the potential benefit of the test. We conclude by providing guidance on appropriate methods for analysis of EHR data informed by our simulation results and data analysis.

2 Methods

2.1 Overview

Let Y represent a binary outcome measure, such as late stage cancer, and S a categorical exposure of interest, such as test indication, with M classes. We assume S is multinomially distributed. In regression models, we convert S into a vector of $M-1$ dummy variables, $U = (U_1, \dots, U_{M-1})^T$, where $U_k = 1$ denotes membership in the $(k+1)$ th class and $U_k = 0$ for all k indicates membership in the first class. Let p denote the vector of probabilities for U , for instance, derived from an externally validated algorithm for predicting the probability of class membership. In the example of test indication where the exposure has only two classes, screening and diagnostic, U is a scalar binary variable taking the value 1 with probability p . Observed covariates, Z , may also be available. We assume that the expectation of Y is related to U and Z via some function of the linear predictor with regression parameters, β . In numerical examples we focus on the case of the logistic link function and define $E(Y|U, Z) = f(U^T, Z^T) \beta = \exp((U^T, Z^T) \beta) / (1 + \exp((U^T, Z^T) \beta))$. Below we describe four approaches to estimating β when U is unknown but p is available via an existing, validated prediction model.

2.2 Substitution

A simple approach to analyzing an imperfectly ascertained exposure is to regress the outcome on $\mathbf{p}$ rather than unknown $\mathbf{U}$. This is an example of Berkson measurement error because unobserved $\mathbf{U}$ is replaced by its expectation. Let $\boldsymbol{\beta}^*$ denote regression model parameters for the regression using the predicted probabilities as exposures in place of the true exposure. In general, since $E(\mathcal{K}(\mathbf{U}^T, \mathbf{Z}^T) \boldsymbol{\beta})$ is not equal to $\mathcal{K}(E(\mathbf{U}^T), \mathbf{Z}^T) \boldsymbol{\beta}$, $\boldsymbol{\beta}^*$ is not equal to $\boldsymbol{\beta}$ so that simply incorporating $\mathbf{p}$ as a predictor in a non-linear regression model results in biased estimates of association.

2.3 Classification

Another simple approach is to dichotomize the predicted probability at some pre-specified cutpoint to create a proxy for the categorical predictor of interest and then regress the outcome on this proxy. The specific threshold for classification could be chosen to obtain desired operating characteristics. Some existing algorithms that provide predicted probabilities also report sensitivity and specificity for correct classification corresponding to a given classification threshold. For instance, the algorithm for colonoscopy indication developed by Adams et al. (2015) had sensitivity of 88.5% and specificity of 90.5% when predicted probabilities were dichotomized at 0.261. Although straightforward to implement, creating a categorical proxy for $\mathbf{U}$ based on $\mathbf{p}$ does not account for the error in this imperfect proxy.

2.4 Mixture model

Information about the nature of the measurement error must be incorporated into the analysis in order to obtain unbiased estimates for $\boldsymbol{\beta}$. This can be accomplished by using maximum likelihood estimation based on the full likelihood incorporating both the probability of class membership and the probability of the outcome conditional on class membership via a mixture distribution. We write the mixture likelihood incorporating these two components as

$$L_{Y|\mathbf{p}}(\boldsymbol{\beta}) = \sum_{k=1}^M p_k f(\mathbf{V}_k^T \boldsymbol{\beta})^Y (1 - f(\mathbf{V}_k^T \boldsymbol{\beta}))^{1-Y},$$

where $\mathbf{V}_k = (\mathbf{U}_k, \mathbf{Z}^T)^T$ is the vector of covariates that would result for a member of class k and p_k is the predicted probability of membership in class k . This likelihood is a mixture of Bernoullis with $\mathbf{p}$ giving the mixing proportions. Maximum likelihood estimates for $\boldsymbol{\beta}$ based on the full likelihood are available via the EM algorithm for a generalized linear finite mixture model (GLFMM) (Jansen 1993).

Specifically, we can write the score function for the mixture likelihood as

$$S_{Y|\mathbf{p}}(\boldsymbol{\beta}) = \sum_{k=1}^M \frac{p_k f(\mathbf{V}_k^T \boldsymbol{\beta})^Y (1 - f(\mathbf{V}_k^T \boldsymbol{\beta}))^{1-Y}}{\sum_{j=1}^M p_j f(\mathbf{V}_j^T \boldsymbol{\beta})^Y (1 - f(\mathbf{V}_j^T \boldsymbol{\beta}))^{1-Y}} \frac{\partial}{\partial \boldsymbol{\beta}} \log [f(\mathbf{V}_k^T \boldsymbol{\beta})^Y (1 - f(\mathbf{V}_k^T \boldsymbol{\beta}))^{1-Y}].$$

This corresponds to the standard score function for a Bernoulli distributed random variable if V_k were known and each observation were replicated M times and weighted by

$$w_k = \frac{p_k f(\mathbf{V}_k^T \boldsymbol{\beta})^Y (1 - f(\mathbf{V}_k^T \boldsymbol{\beta}))^{1-Y}}{\sum_{j=1}^M p_j f(\mathbf{V}_j^T \boldsymbol{\beta})^Y (1 - f(\mathbf{V}_j^T \boldsymbol{\beta}))^{1-Y}}. \text{ This facilitates straightforward implementation of the EM algorithm using standard software for GLMs via the following approach:}$$

1. Replicate each observation in the data set M times.
2. (E-step) Assign each replicate a weight, w_k , using the current value of $\hat{\boldsymbol{\beta}}$ to estimate $f(\mathbf{V}_k^T \boldsymbol{\beta})$. At the first iteration, $\hat{\boldsymbol{\beta}}^*$ can be used as an approximation for $\hat{\boldsymbol{\beta}}$.
3. (M-step) Obtain updated estimates by fitting a weighted GLM to the pseudo-data with weights $\mathbf{w} = w_1, \dots, w_M$.

Repeat steps 2 and 3 until the algorithm reaches convergence.

In addition to its relatively simple implementation, the ML estimates for the mixture model parameters returned by the EM algorithm are consistent and asymptotically normally distributed (Redner, Walker 1984).

2.5 Multiple imputation

There are many similarities between measurement error and missing data problems. Above we considered $\mathbf{p}$ to be a surrogate for $\mathbf{U}$ that was measured with error. Alternatively, we can consider $\mathbf{U}$ to represent missing data, with additional information about $\mathbf{U}$ available via $\mathbf{p}$. While the measurement error conceptualization of the problem leads directly to the mixture model presented above, the missing data formulation suggests an approach via multiple imputation (MI).

This approach is appealing given the availability of standard software for MI and a potential imputation model, $\mathbf{p}$. In latent class analyses, the MI approach is frequently used for estimating the association between latent class membership and outcomes, but has been demonstrated to induce bias (Vermunt 2010). Bias in $\hat{\boldsymbol{\beta}}$ arises because information on the relationship between $\mathbf{U}$ and $\mathbf{Y}$ has not been incorporated into the imputation model. Unlike standard missing data problems, the exposure is missing for everyone so it is not possible to directly impute the exposure conditional on the observed data. A straightforward alternative to imputing $\mathbf{U}$ including information from both $\mathbf{p}$ and $\mathbf{Y}$ is to simulate from the distribution of $\mathbf{U} | \mathbf{p}, \mathbf{Y}$, which follows a multinomial distribution with mean $\mathbf{w}$, as given in step 2 of the estimation algorithm described above. Since $\mathbf{w}$ depends on $\boldsymbol{\beta}$, an approximate value for $\mathbf{w}$ can be obtained by substituting $\boldsymbol{\beta}^*$ for $\boldsymbol{\beta}$. Values for $\mathbf{U}$ can be imputed from this distribution and results combined across iterations using standard rules for multiple imputation (Little, Rubin 2002). Typically, likelihood-based estimation using the EM algorithm and estimation based on MI return comparable results. However, in this case results will differ because the approximation, $\boldsymbol{\beta}^*$, is used in the MI approach, whereas in the full likelihood approach this approximation is only used in the first iteration and is subsequently updated.

2.6 Simulation study

We conducted simulation studies to compare the bias and efficiency of the four alternative approaches to estimating the association between an imperfectly ascertained exposure and an outcome. We simulated data for samples of 10,000 individuals. The choice of sample size for the simulation studies was motivated by EHR-based studies, which typically include several thousand individuals (e.g., (Levin et al. 2006; Tamblyn et al. 2014; Wu et al. 2011)). In simulations, we assumed S was dichotomous and, for each individual, simulated a value for p from a Beta density with mean 0.25. The Beta density was chosen because it allowed us to vary the classification accuracy of p . The Beta density can be bimodal and concentrated around 0 and 1, representing a prediction algorithm with high classification accuracy, or unimodal and concentrated around the mean, representing a prediction algorithm with poor classification accuracy. Values for (a, b) ranging from (0.1, 0.3) to (1, 3) correspond to a range of prediction models with area under the ROC curve similar to those that have been achieved by algorithms for assigning colonoscopy indication (El-Serag et al. 2006; Fisher et al. 2011; Adams et al. 2015), with (0.1, 0.3) representing a prediction model with greater classification accuracy and (1, 3) representing a prediction model with poorer classification accuracy. We therefore investigated these two sets of values in simulation studies.

Next, we simulated a value for the dummy variable indicating the true (unobserved) classification, U , for each individual that was drawn from a Bernoulli distribution with mean p . Finally, for each individual we simulated a value for the outcome of interest, Y , drawing values from a Bernoulli distribution with mean $\exp(\beta_0 + \beta_1 U) / (1 + \exp(\beta_0 + \beta_1 U))$. We selected values for β_0 corresponding to common and rare outcomes (prevalence of 0.3 and 0.1 in the unexposed group) and evaluated values for β_1 spanning a range from no association to large odds ratios (OR) associated with the exposure (log OR ranging from 0–1.4 corresponding to OR ranging from 1 – 4).

For the full likelihood approach, the EM algorithm was iterated until parameter estimates from successive iterations differed by less than 10^{-9} . This convergence criterion was chosen because 10^{-9} was considered to be beyond the limit of precision that might reasonably be of interest in a scientific study.

For the MI approach we first estimated $\hat{\beta}^*$ and its standard error by regressing Y on p . We then simulated a random value from a multivariate normal distribution with mean $\hat{\beta}^*$ and variance given by the estimated variance of $\hat{\beta}^*$. Finally, we imputed values for U by sampling from a Bernoulli distribution with mean w , substituting our draw from the distribution of $\hat{\beta}^*$ for β . The extra step of simulating a value from the sampling distribution of $\hat{\beta}^*$ was included to reflect variability due to the uncertainty in our estimate of $\hat{\beta}^*$ as well as uncertainty in U . Each MI estimate was based on five imputed data sets. Each simulation scenario was repeated for 1,000 iterations. We computer bias and empirical standard errors across the 1,000 iterations.

2.7 Group Health colorectal cancer data

To demonstrate how the choice of analytic approach can influence study results, we applied all four approaches to an example study where test indication was the exposure. In this analysis we investigated the association between colonoscopy indication and late stage cancer diagnosis. The objective of this analysis was to compare results of alternative methods for estimating this association in the presence of imperfect information on the exposure, test indication. We used data from Group Health, a large integrated health care system in Washington State serving approximately 600,000 members, to investigate this hypothesis by estimating the odds ratio associated with detecting a late stage cancer at screening compared to diagnostic colonoscopy. If screening colonoscopy were effective we would expect a smaller proportion of cancers diagnosed on screening examinations to be late stage cancers compared to those diagnosed at diagnostic examinations (Winawer et al. 1997). Data from the Group Health EHR were available for patient age, race, and utilization of colonoscopy and other medical procedures. Colorectal cancer diagnoses were ascertained by linkage to the Puget Sound Surveillance Epidemiology and End Results (SEER) program. This project was approved by the Group Health Institutional Review Board.

We used information on colorectal cancers diagnosed within 60 days of a colonoscopy performed on individuals aged 50–79 years between January 1, 1993 and December 31, 2011. A cancer diagnosed within 60 days of colonoscopy was considered to have been “detected” by that colonoscopy. In Group Health data, 89% of colorectal cancers are preceded by a colonoscopy in the prior 60 days. Due to the high sensitivity of colonoscopy and slow growth of colorectal cancer, it is highly unlikely that this definition would result in inclusion of colonoscopies that were unrelated to the subsequent cancer diagnosis. Each colonoscopy was assigned a probability of having a test indication of “screening” using a previously developed algorithm based on International Classification of Diseases, version 9 and Common Procedural Terminology codes for conditions and procedures that indicate signs or symptoms of colorectal cancer (Adams et al. 2015). In this prior work, the cut point in the probability of screening test indication that maximized the sum of sensitivity and specificity was 0.261. We used this cut point to dichotomize probabilities for the classification approach. Late stage cancer was defined as a cancer with American Joint Committee on Cancer (AJCC) version 6 stage 2B or higher.

We analyzed the association between screening test indication and late stage at diagnosis using the four methods described in Section 2.6. In each model, we adjusted for age and race, as these covariates were available in EHR data and are strongly associated with colorectal cancer screening and cancer characteristics (Siegel et al. 2014). In addition to possible misclassification of test indication, other data elements derived from the EHR may also contain errors. However, we anticipate that data required for clinical care or billing are likely to be of high quality. For this reason, age is likely measured with good accuracy. Similarly, data on colonoscopy utilization are based on procedure codes required for billing, making it likely that these procedures were performed. Of the variables included in our analysis, only indication and race are expected to contain errors. Previous studies of concordance between self-reported and EHR-derived race have found modest agreement

(Boehmer et al. 2002; Gomez et al. 2005; West et al. 2005). Misclassification of race could result in imperfect adjustment for confounding.

3 Results

3.1 Simulation results

Bias in the substitution, classification, and multiple imputation approaches tended to increase with increasing log OR (Figure 1). This trend was stronger for rarer outcomes (prevalence = 0.1) and when predicted probabilities had poorer discriminatory accuracy ($a = 1$, $b = 3$). The mixture model was nearly unbiased across all scenarios investigated. The MI approach displayed bias generally intermediate between that of the substitution approach and the mixture model. The classification approach was extremely biased across all scenarios investigated. In Figure 1, we have truncated the y-axis at -0.1 in order to facilitate comparisons across the mixture, MI, and substitution approaches. Bias in the classification approach exceeded -0.3 at $\log OR = 1.4$ when $a = 0.1$ and $b = 0.3$, and exceeded -0.8 at $\log OR = 1.4$ when $a = 1$ and $b = 3$.

In Figure 2, we compared empirical standard errors across the four approaches at prevalence = 0.1 and $\log OR = 0.69$. Trends in relative efficiency of the methods were the same for other values of prevalence and strength of association. In the case of greater discriminatory accuracy ($a = 0.1$, $b = 0.3$), all four methods were similarly efficient. In the case of poorer discriminatory accuracy ($a = 1$, $b = 3$), the substitution method displayed larger standard errors than the other three approaches. Although the classification approach was similarly efficient to other methods, bias in this method is notably larger.

3.2 Results for Group Health colorectal cancer data

To estimate the association between screening colonoscopy and colorectal cancer stage at diagnosis, we identified 843 colorectal cancer diagnoses meeting inclusion criteria. The mean algorithm-assigned probability of test indication of screening was 16.0%. In this cohort, 306 (36.3%) individuals were diagnosed with a late stage colorectal cancer. Patient demographics differed somewhat between those diagnosed with a late stage cancer and those with an earlier stage cancer (Table 1). A higher proportion of patients with late stage cancer were younger and non-white compared to patients whose cancer was diagnosed at an earlier stage.

All four methods indicated a significant decrease in the odds of late stage diagnosis for cancers diagnosed at screening colonoscopy compared to diagnostic colonoscopy after adjustment for age and race/ethnicity (Table 2). Point estimates from the substitution and MI approaches were relatively similar, while the classification approach was substantially attenuated towards the null and the mixture distribution approach indicated a larger decrease in odds of late stage cancer associated with screening colonoscopy compared to other methods. This pattern corresponds to the findings of our simulation studies where the classification approach was strongly biased towards the null while the missing data and substitution approaches displayed relatively less bias. Based on results of simulation studies,

we anticipate that the mixture model result is the most reliable in this application because it displayed little or no bias across all scenarios investigated.

4 Discussion

When a categorical exposure variable is unknown but predicted probabilities of class membership for that exposure are available, regression parameter estimates based on the full likelihood constructed using a mixture distribution are unbiased and can be relatively easily obtained using the EM algorithm. By formulating the likelihood as a mixture distribution with known mixing probabilities based on predicted probabilities of class membership, estimates can be obtained via standard software for generalized linear models. An alternative missing data formulation also decreases bias relative to the naïve approach of using the predicted probabilities as exposure variables. However, because parameters of the distribution of the exposure conditional on both predicted probabilities and the outcome are unknown and can only be approximated, this method will decrease bias but not eliminate it. In general, the MI approach will not fully correct the bias attributable to unknown exposure class membership and is not recommended.

Previous work has shown that the approach of replacing an unknown categorical exposure variable with its predicted probability is unbiased in linear regression models and suffers from minimal loss of efficiency (McCaffrey, Elliott 2008). In our simulations using a logistic link function, we found that the substitution approach was minimally biased for common outcomes, especially when predicted probabilities with good discriminatory accuracy were available and when the true odds ratio was small to moderate. This is likely because the logistic curve is approximately linear over small regions, especially at higher prevalence. This approach to estimating associations between imperfectly ascertained categorical exposures and outcomes by substituting predicted probabilities would be reasonable in settings with common outcomes and where moderate or smaller odds ratios are anticipated.

Stage shift in screen-detected cancers, the hypothesized decrease in cancer stage among cases diagnosed at screening, is one example of a research question that can be investigated using EHR data in which exposure class membership is not definitively known. Because colorectal cancer, and particularly late stage colorectal cancer, is a rare outcome, large samples such as those available from EHRs are needed in order to identify a sufficiently large pool of cases in which to explore this relationship. However, such data sources often lack information on indication for colonoscopy. Previous observational studies have identified a significant decrease in colorectal cancer mortality for individuals screened with colonoscopy (Lin et al. 2014; Brenner et al. 2014). Our analysis builds on this prior work by demonstrating that screening colonoscopy is associated with earlier stage at cancer diagnosis, a possible mechanism by which screening may affect mortality. In our analysis of stage-shift in colorectal cancer we found that dichotomizing predicted probabilities for screening test indication resulted in substantial attenuation of parameter estimates, consistent with our findings in simulation studies.

The problem of unknown class membership for exposures of interest is a common feature in research using EHR data. Numerous “computable phenotypes,” EHR-based algorithms

providing predicted probabilities for conditions of interest, have recently been developed (Sun et al. 2014; Ananthakrishnan et al. 2013; Liao et al. 2010). Our research suggests a straightforward, unbiased method for incorporating predicted probabilities from these models as exposures in analyses using GLMs.

Our study provides guidance on a practical approach to using predicted probabilities of exposures in analyses of data from EHRs. The mixture distribution approach corrects bias with no loss of efficiency across the range of odds ratios, prevalences, and predictive model performances investigated in this study and can be implemented easily using standard software for GLMs.

Acknowledgments

This work was supported by the National Cancer Institute of the National Institutes of Health (grant number U01CA152959). The collection of cancer incidence data used in this study was supported by the Cancer Surveillance System of the Fred Hutchinson Cancer Research Center, which is funded by Contract No. N01-CN-67009 and N01-PC-35142 from the Surveillance, Epidemiology and End Results (SEER) Program of the National Cancer Institute with additional support from the Fred Hutchinson Cancer Research Center and the State of Washington.

References

- Adams KF, Johnson EA, Chubak J, Kamineni A, Doubeni CA, Buist DSM, Weinmann S, Doria-Rose VP, Rutter CM. Ascertainment of colonoscopy indication using administrative data. *Egms*. 2015; 3(1):11.
- Ananthakrishnan AN, Cai T, Savova G, Cheng SC, Chen P, Perez RG, Gainer VS, Murphy SN, Szolovits P, Xia Z, Shaw S, Churchill S, Karlson EW, Kohane I, Plenge RM, Liao KP. Improving case definition of Crohn's disease and ulcerative colitis in electronic medical records using natural language processing: a novel informatics approach. *Inflammatory bowel diseases*. 2013; 19(7): 1411–1420. [PubMed: 23567779]
- Boehmer U, Kressin NR, Berlowitz DR, Christiansen CL, Kazis LE, Jones JA. Self-reported vs administrative race/ethnicity data and study results. *American journal of public health*. 2002; 92(9): 1471–1473. doi: [PubMed: 12197976]
- Brenner H, Stock C, Hoffmeister M. Effect of screening sigmoidoscopy and screening colonoscopy on colorectal cancer incidence and mortality: systematic review and meta-analysis of randomised controlled trials and observational studies. *Bmj*. 2014; 348:g2467. [PubMed: 24922745]
- El-Serag HB, Petersen L, Hampel H, Richardson P, Cooper G. The use of screening colonoscopy for patients cared for by the Department of Veterans Affairs. *Arch Intern Med*. 2006; 166(20):2202–2208. [PubMed: 17101937]
- Elliott MN, Morrison PA, Fremont A, McCaffrey DF, Pantoja P, Lurie N. Using the census bureau's surname list to improve estimates of race/ethnicity and associated disparities. *Health Services and Outcomes Research Methodology*. 2009; 9(2):69–83.
- Fisher DA, Grubber JM, Castor JM, Coffman CJ. Ascertainment of colonoscopy indication using administrative data. *Dig Dis Sci*. 2011; 55(6):1721–1725.
- Gomez SL, Kelsey JL, Glaser SL, Lee MM, Sidney S. Inconsistencies between self-reported ethnicity and ethnicity recorded in a health maintenance organization. *Annals of Epidemiology*. 2005; 15(1): 71–79. [PubMed: 15571996]
- Hardcastle JD, Chamberlain JO, Robinson MH, Moss SM, Amar SS, Balfour TW, James PD, Mangham CM. Randomised controlled trial of faecal-occult-blood screening for colorectal cancer. *Lancet*. 1996; 348(9040):1472–1477. [PubMed: 8942775]
- Hunt L, Jorgensen M. Mixture model clustering for mixed data with missing information. *Comput Stat Data An*. 2003; 41(3–4):429–440. doi:Pii S0167-9473(02)00190-1.

- Jansen R. Maximum likelihood in a generalized linear finite mixture model by using the EM algorithm. *Biometrics*. 1993; 49:227–231.
- Kronborg O, Fenger C, Olsen J, Jorgensen OD, Sondergaard O. Randomised study of screening for colorectal cancer with faecal-occult-blood test. *Lancet*. 1996; 348(9040):1467–1471. [PubMed: 8942774]
- Levin TR, Zhao W, Conell C, Seeff LC, Manninen DL, Shapiro JA, Schulman J. Complications of colonoscopy in an integrated health care delivery system. *Annals of internal medicine*. 2006; 145(12):880–886. [PubMed: 17179057]
- Liao KP, Cai T, Gainer V, Goryachev S, Zeng-treitler Q, Raychaudhuri S, Szolovits P, Churchill S, Murphy S, Kohane I, Karlson EW, Plenge RM. Electronic medical records for discovery research in rheumatoid arthritis. *Arthritis care & research*. 2010; 62(8):1120–1127. [PubMed: 20235204]
- Lin OS, Kozarek RA, Cha JM. Impact of sigmoidoscopy and colonoscopy on colorectal cancer incidence and mortality: an evidence-based review of published prospective and retrospective studies. *Intestinal research*. 2014; 12(4):268–274. [PubMed: 25374491]
- Little, RJ., Rubin, DB. *Statistical analysis with missing data*. 2. Hoboken, NJ: Wiley; 2002.
- Mandel JS, Bond JH, Church TR, Snover DC, Bradley GM, Schuman LM, Ederer F. Reducing mortality from colorectal cancer by screening for fecal occult blood. Minnesota Colon Cancer Control Study. *N Engl J Med*. 1993; 328(19):1365–1371. [PubMed: 8474513]
- McCaffrey DF, Elliott MN. Power of tests for a dichotomous independent variable measured with error. *Health Serv Res*. 2008; 43(3):1085–1101. [PubMed: 18454782]
- Redner RA, Walker HF. Mixture Densities, Maximum-Likelihood and the Em Algorithm. *Siam Rev*. 1984; 26(2):195–237. doi:
- Richesson R, Smerek M. Electronic health records-based phenotyping. 2015 [Accessed November 30, 2015] <http://sites.duke.edu/rethinkingclinicaltrials/informed-consent-in-pragmatic-clinical-trials/>.
- Siegel R, Desantis C, Jemal A. Colorectal cancer statistics, 2014. *CA: a cancer journal for clinicians*. 2014; 64(2):104–117. [PubMed: 24639052]
- Sun JM, McNaughton CD, Zhang P, Perer A, Gkoulalas-Divanis A, Denny JC, Kirby J, Lasko T, Saip A, Malin BA. Predicting changes in hypertension control using electronic health records from a chronic disease management program. *Journal of the American Medical Informatics Association : JAMIA*. 2014; 21(2):337–344. doi: [PubMed: 24045907]
- Tabar L, Vitak B, Chen HH, Duffy SW, Yen MF, Chiang CF, Krusemo UB, Tot T, Smith RA. The Swedish Two-County Trial twenty years later. Updated mortality results and new insights from long-term follow-up. *Radiol Clin North Am*. 2000; 38(4):625–651. [PubMed: 10943268]
- Tamblyn R, Egale T, Huang A, Winslade N, Doran P. The incidence and determinants of primary nonadherence with prescribed medication in primary care: a cohort study. *Annals of internal medicine*. 2014; 160(7):441–450. [PubMed: 24687067]
- Thompson TJ, Smith PJ, Boyle JP. Finite mixture models with concomitant information: assessing diagnostic criteria for diabetes. *Journal of the Royal Statistical Society Series C-Applied Statistics*. 1998; 47:393–404.
- US Preventive Services Task Force. Screening for colorectal cancer: U.S. Preventive Services Task Force recommendation statement. *Ann Intern Med*. 2008; 149(9):627–637. doi: 0000605-200811040-00243 [pii]. [PubMed: 18838716]
- Vermunt JK. Latent Class Modeling with Covariates: Two Improved Three-Step Approaches. *Polit Anal*. 2010; 18(4):450–469.
- Weiss NS. Analysis of case-control studies of the efficacy of screening for cancer: How should we deal with tests done in persons with symptoms? *American journal of epidemiology*. 1998; 147(12): 1099–1102. [PubMed: 9645787]
- Weiss NS, McKnight B, Stevens NG. Approaches to the analysis of case-control studies of the efficacy of screening for cancer. *American journal of epidemiology*. 1992; 135(7):817–823. [PubMed: 1595681]
- West CN, Geiger AM, Greene SM, Harris EL, Liu IL, Barton MB, Elmore JG, Rolnick S, Nekhlyudov L, Altschuler A, Herrinton LJ, Fletcher SW, Emmons KM. Race and ethnicity: comparing medical records to self-reports. *Journal of the National Cancer Institute. Monographs*. 2005; (35):72–74.

- Winawer SJ, Fletcher RH, Miller L, Godlee F, Stolar MH, Mulrow CD, Woolf SH, Glick SN, Ganiats TG, Bond JH, Rosen L, Zapka JG, Olsen SJ, Giardiello FM, Sisk JE, vanAntwerp R, BrownDavis C, Marciniak DA, Mayer RJ. Colorectal cancer screening: Clinical guidelines and rationale. *Gastroenterology*. 1997; 112(2):594–642. doi: [PubMed: 9024315]
- Wu LT, Gersing K, Burchett B, Woody GE, Blazer DG. Substance use disorders and comorbid Axis I and II psychiatric disorders among young psychiatric patients: Findings from a large electronic health records database. *J Psychiatr Res*. 2011; 45(11):1453–1462. [PubMed: 21742345]

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Figure 1.

Bias in estimated log odds ratio (OR) across a range of true values for log OR for four methods for estimating associations under exposure misclassification. Dotted line indicates no bias. Black = substitution, red = classification, green = mixture model, blue = multiple imputation.

Figure 2.

Estimated log odds ratios (ORs) with empirical standard error bars for four methods for estimating associations under exposure misclassification. Dotted line indicates the true log OR value of 0.69. Subs = substitution, Cut = classification, Mix = mixture model, MI = multiple imputation.

Table 1

Characteristics of 843 individuals from Group Health with a diagnosis of colorectal cancer stratified by late stage disease detected at colonoscopy

	Early stage N (%)	Late stage N (%)
Age (years)		
50–59	132 (24.6)	89 (29.1)
60–69	172 (32.0)	105 (34.3)
70–79	233 (43.4)	112 (36.6)
Race/Ethnicity		
Non-Hispanic White	477 (88.8)	267 (87.3)
Non-Hispanic Black	16 (3.0)	14 (4.6)
Asian/Pacific Islander	30 (5.6)	15 (4.9)
Hispanic	10 (1.9)	8 (2.6)
Native American	4 (0.7)	2 (0.7)

Table 2

Odds ratios (OR), 95% confidence intervals (CI) and P-values for odds of a late stage cancer detected at screening compared to diagnostic colonoscopy after adjustment for age and race/ethnicity

	OR	95% CI	P-value
Substitution	0.23	(0.08, 0.63)	0.005
Classification	0.60	(0.39, 0.89)	0.013
Mixture distribution	0.10	(0.02, 0.62)	0.013
MI	0.22	(0.13, 0.37)	<0.001

Abbreviations: MI = multiple imputation