

RESEARCH ARTICLE

Open Access



Accounting for misclassification bias of binary outcomes due to underscreening: a sensitivity analysis

Nanhua Zhang^{1,2*}, Si Cheng¹, Lilliam Ambroggio^{1,2}, Todd A. Florin^{2,3} and Maurizio Macaluso^{1,2}

Abstract

Background: Diagnostic tests are performed in a subset of the population who are at higher risk, resulting in undiagnosed cases among those who do not receive the test. This poses a challenge for estimating the prevalence of the disease in the study population, and also for studying the risk factors for the disease.

Methods: We formulate this problem as a missing data problem because the disease status is unknown for those who do not receive the test. We propose a Bayesian selection model which models the joint distribution of the disease outcome and whether testing was received. The sensitivity analysis allows us to assess how the association of the risk factors with the disease outcome as well as the disease prevalence change with the sensitivity parameter.

Results: We illustrated our model using a retrospective cohort study of children with asthma exacerbation that were evaluated for pneumonia in the emergency department. Our model found that female gender, having fever during ED or at triage, and having severe hypoxia are significantly associated with having radiographic pneumonia. In addition, simulation studies demonstrate that the Bayesian selection model works well even under circumstances when both the disease prevalence and the screening proportion is low.

Conclusion: The Bayesian selection model is a viable tool to consider for estimating the disease prevalence and in studying risk factors of the disease, when only a subset of the target population receive the test.

Keywords: Misclassification, Selection model, Underscreening, Radiographic pneumonia

Background

It is often of interest to estimate the prevalence of a certain condition in a population and this proportion is underestimated as those who were not screened for the disease were assumed negative for the condition. Underscreening of a disease is common because of resource limitation, perceived low risk which eliminates the need of screening, or lack of recommendations from guidelines. For example, fatty liver disease and metabolic syndrome among children may go undetected because relevant evaluations were not routinely recommended by pediatricians

[1]. At patient level, under-diagnosis could cause delay in treatment. A study in Italy reported the rate of Chronic Obstructive Pulmonary Disease (COPD) under-diagnosis ranges between 25 and 50%, and as a consequence, many patients missed the optimal time for therapeutic intervention, contributing to the progression of the disease to be more severe [2]. Other common diseases that go under-diagnosed include hepatitis C virus (HCV) [3, 4], HIV and sexually transmitted diseases (STD) [5], hypertension in children and adolescents [6], and depression [7]. Under-diagnosis of infectious disease such as pneumonia, HIV or STD poses a societal burden.

Perceived as at low risk for the condition, those who are not screened or examined are usually classified as negative, which results in an underestimated proportion or prevalence of the disease in the study population.

* Correspondence: nanhua.zhang@cchmc.org

¹Division of Biostatistics and Epidemiology, Cincinnati Children's Hospital Medical Center, 3333 Burnet Ave, MLC 5041, Cincinnati, OH 45229, USA

²Department of Pediatrics, University of Cincinnati College of Medicine, Cincinnati, OH, USA

Full list of author information is available at the end of the article



This misclassification can also bias the association of risk factors with the disease condition [8–10]. Standard approach to handle misclassification in binary outcomes relies on validation study of a subsample of initial non-respondents in the study population. When validation data are not available, Hausman et al. (1998) [11] and Savoca (2011) [12] examine the misclassification bias as a function of the error rates, under balanced and unbalanced scenarios. However, these methods require assumptions on the functional form of the positive diagnosis probability and the misclassification parameters [12]. Shebl et al. (2012) develop a likelihood-based method to estimate incidence when disease status is measured imperfectly based on hidden Markov models, while assuming the known constant levels of sensitivity and specificity and constant incidence rates over time [4].

Technically, the disease condition for those who were not screened/examined is unknown and classifying them as negative is based on a strong assumption about the missing values. Instead, we formulate the problem as a missing data problem and treat the disease status of those not tested as missing. The missing data mechanism, which concerns how the data are missing and whether the missingness is related to the underlying missing values, is critical when dealing with missingness. When the missingness depends neither on observed nor missing values, the data are missing completely at random (MCAR). When the missingness depends on the observed values but not the missing values, the missing data mechanism is called missing at random (MAR). In the case when the missingness can depend on the missing values, the missing data mechanism is called missing not at random (MNAR). When the missing data mechanism is MCAR or MAR, correct inference may be achieved based on a likelihood function which does not involve a modeling for the missingness mechanism; likelihood inference which ignores the model for missingness (ignorable likelihood, Zhang and Little 2011 [13]) includes maximum likelihood estimation, Bayesian inference, and multiple imputation [14]. However, when the missing data mechanism is MNAR, a correct inference has to consider the joint distribution of the outcome variable and the missingness indicator; depending on factorization of the joint distribution of the outcome variable and the missingness indicator, three classes of models have been investigated: the selection model, pattern mixture model, and shared parameter model [15].

In this article, we propose a class of Bayesian selection model, which estimates the disease prevalence in the study population (both screened and unscreened) using a sensitivity parameter which denotes the likelihood of

being screened. This model will yield estimates of the prevalence as well as the association of risk factors with the disease outcome under different values of the sensitivity parameter and therefore the big picture of the research questions.

Methods

Data source

We were interested in estimating the proportion of children with an asthma exacerbation who were diagnosed with radiographic pneumonia among those presented to the emergency department of the Cincinnati Children's Hospital Medical Center between January 1st, 2010, and December 31, 2013. Children were identified using a validated algorithm of *an International Classification of Diseases, Ninth Revision, Clinical Modification* diagnosis code of asthma (code 493.x) in the first 3 diagnosis positions and receipt of 1 or more doses of albuterol sulfate in the emergency department [16]. Children less than 2 years were excluded to minimize including infants with bronchiolitis.

We investigated the risk factors for radiographic pneumonia i.e. focal opacity present on chest radiograph) [17]. Consequences of the overuse of radiography include increased time in the hospital, unnecessary radiation, increased cost, and inappropriate antibiotic use due to equivocal imaging findings [18]. Due to the high rate of normal chest radiograph and the consequences of unnecessary radiograph, only about a third of those who presented to emergency department received chest radiography. This was noted as a limitation of the regression analysis used as only subjects who received chest radiography were included in fitting the model that assessed risk factors for radiographic pneumonia and therefore limited the generalizability of their findings to the larger study population of all children with asthma exacerbation who present to the ED [17]. Due to the fact that those who received chest radiography are a biased sample of all presented to ED with asthma exacerbation, with possibly higher probability of having radiographic pneumonia than children who did not receive the chest radiography, the analyses that discarded the subjects who did not receive the chest radiography may have led to biased estimation of the risk factors with the outcome of radiographic pneumonia.

Assuming those who did not undergo chest radiography to be negative for radiographic pneumonia will underestimate the prevalence of the disease in the study population and potentially bias the association of risk factors on the disease. We formulate this problem as a missing problem and use a Bayesian selection model to jointly model the disease status and the response indicator, i.e., whether the subject received chest radiography or not.

Bayesian selection model

Let y_i denote the actual binary radiographic pneumonia status, equal to 1 if the i^{th} subject had radiographic pneumonia, and 0 if not. This outcome was observed for subjects who received chest radiography, and missing for subjects who did not receive chest radiography. We use R_i to denote whether we observed the i^{th} subjects radiographic pneumonia status, and R_i is equal to 1 if the subject received chest radiography and equal to 0 if not. We use x_i and z_i to denote the covariate sets that predicts the outcome y_i and the response status R_i , respectively. The covariates in x_i and z_i may overlap with each other. The selection model is based on the joint distribution of (y_i, R_i) ,

$$f(y_i, R_i | x_i, z_i; \beta, \lambda, \theta) = f(y_i | x_i; \beta) f(R_i | y_i, z_i; \lambda, \theta) \quad (1)$$

where $f(y_i | x_i; \beta)$ and $f(R_i | y_i, z_i; \theta)$ are modeled as logistic regression as

$$\text{logit}(\Pr(y_i = 1 | x_i; \beta)) = x_i^T \beta \quad (2)$$

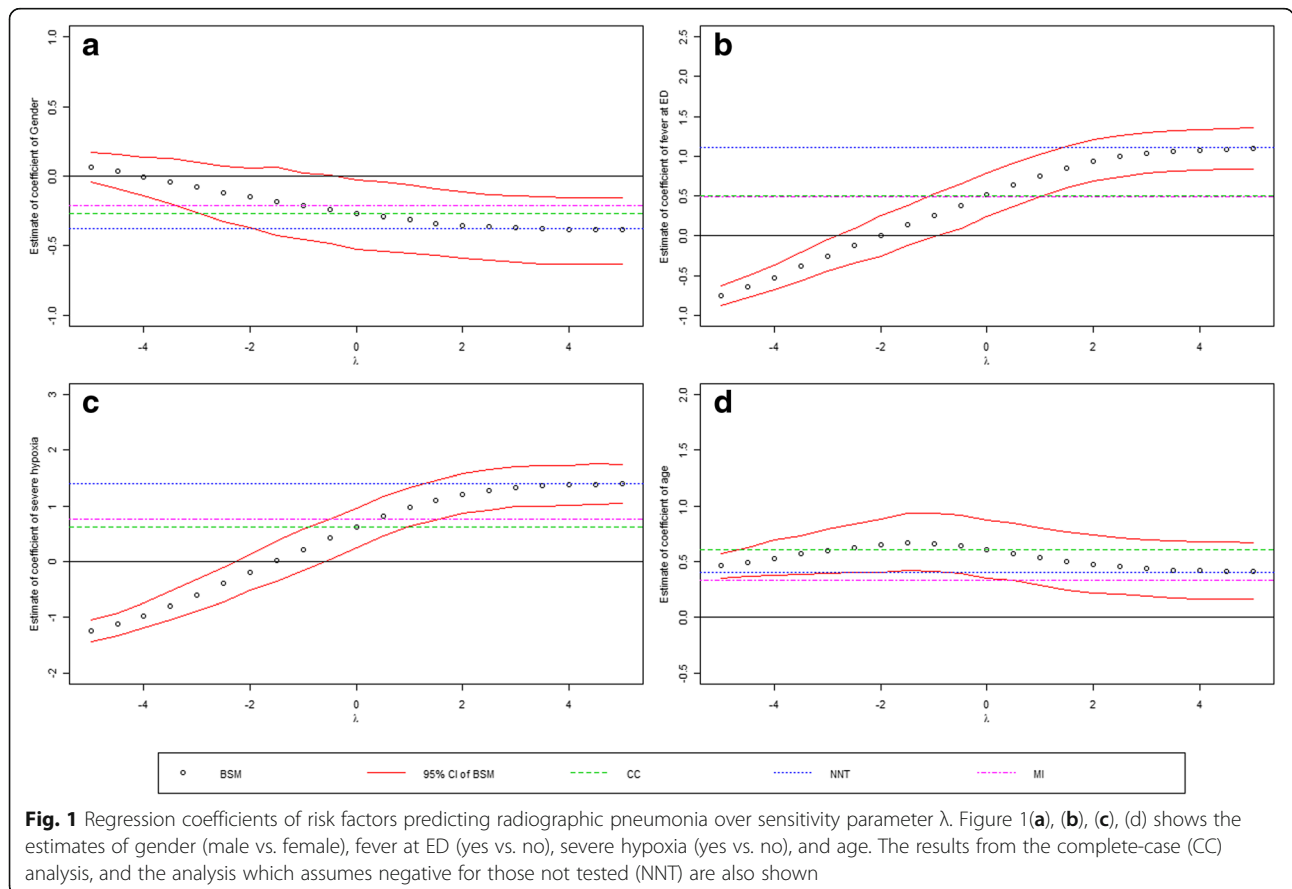
$$\text{logit}(\Pr(R_i = 1 | y_i, z_i; \theta)) = z_i^T \theta + \lambda y_i \quad (3)$$

Here the parameter β denotes the risk for radiographic pneumonia, which is the main parameter of interest; and

the parameters θ and λ relate the propensity of receiving chest radiography (and hence the response indicator) to covariates z_i and the actual pneumonia status y_i . Note here y_i is missing for subjects who did not receive radiography, which leads to identification issues for this joint model [15].

To address the identification issues inherent in the model, we use λ as a sensitivity parameter, taking a range of fixed values from $-\infty$ to ∞ . When λ is 0, the propensity of a subject receiving chest radiography does not depend on this subject's radiographic pneumonia status; this corresponds to missing at random assumption in the missing data literature. When λ is greater than 0, the propensity of a subject receiving chest radiography is positively associated with the subject's radiographic pneumonia status. When λ is less than 0, having radiographic pneumonia is associated with lower propensity of receiving a chest radiography. For this specific application, it is reasonable to assume that patients with radiographic pneumonia are more likely to receive chest radiography than patients without radiographic pneumonia, and therefore $\lambda > 0$.

An important feature of the model is that by allowing the sensitivity parameter λ to change, we can assess how the main parameters of interest is sensitive to the



perturbation of the sensitivity parameter. The Bayesian modeling setup also makes it easy to predict the overall proportion of the radiographic pneumonia in the study population for a fixed λ . For illustration purpose, we use the same set of covariates for x_i and z_i , which includes gender (female vs. male), age at visit (≥ 5 years vs. < 5 years), fever during ED stay or at triage (temperature ≥ 38 °C vs. < 38 °C), and severe hypoxia (Oxygen saturation $< 90\%$ vs. $\geq 90\%$). [17]

We formulate the model in a Bayesian framework (BSM) and estimate the parameters using Markov Chain Monte Carlo (MCMC) methods. The MCMC algorithm is called “data augmentation”. The algorithm iteratively draws the next values of parameters and the unobserved y_i 's from the corresponding posterior distributions of the parameters and the posterior predictive distributions of the unobserved y_i 's. We use proper and non-informative prior distributions for all parameters, i.e., multivariate normal priors with mean 0 and diagonal covariance matrices with a large scale parameter of 10,000 for both β and θ . The software package WinBUGS is used to estimate the posterior distribution of the parameters [19].

Results

Out of the 14,007 children who visited emergency department for asthma exacerbation, chest radiography was performed on 4708 children (33.6%). Radiographic pneumonia was present in 280 of the 4708 children who received chest radiography (5.9%).

Figure 1(a)-(d) shows the regression parameters of gender (β_1 : males vs. female), age at visit (β_2 : ≥ 5 years vs. < 5 years), fever during ED stay or at triage (β_3 : temperature ≥ 38 °C vs. < 38 °C), and severe hypoxia (β_4 : Oxygen saturation $< 90\%$ vs. $\geq 90\%$), respectively. For comparison purpose, the results from the following two naïve methods were also plotted on the same plots:

- Complete-case analysis (CC): logistic regression only includes those had observed radiographic pneumonia status, i.e., those who received chest radiography;
- Negative for not tested (NNT): logistic regression with all subjects which assumes negative radiographic pneumonia for those who did not receive chest radiography.
- Multiple imputation (MI): multiple imputation using chained equation which assumes missing at random.

The point estimates from the four methods and the 95% credible intervals (CI) from the proposed Bayesian selection model were plotted on the same plots. As the sensitivity parameter goes from -4 to 4 , we see a decreasing trend of the risk of having radiographic pneumonia comparing males to females. As we mentioned before, the true λ should be positive because those with pneumonia are believed to be more likely to receive chest radiography; therefore, we focus on the results when λ is positive. When λ is greater than 0, the coefficients is negative for gender from the BSM and the 95% CI does not cover 0,

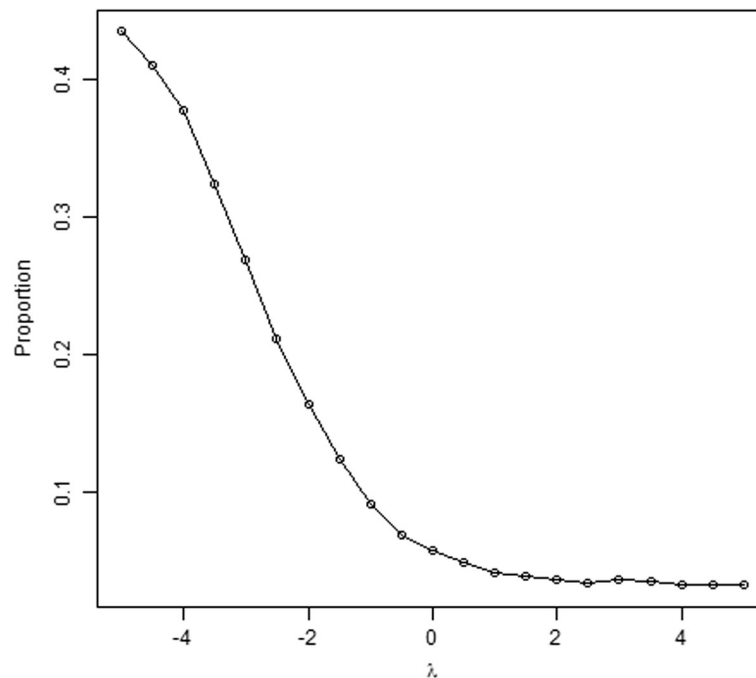


Fig. 2 Estimated prevalence of radiographic pneumonia in the target population by sensitivity parameter λ

implying that males had significantly higher risk of having radiographic pneumonia than females among these who visited emergency department for asthma exacerbation (Fig. 1(a)). For $\lambda > 0$, old age is associated with significant high risk of having radiographic pneumonia among this study population (Fig. 1(b)). Having fever during ED or at triage, or having severe hypoxia, are both significantly associated with positive radiographic pneumonia (Fig. 1(c)-(d)).

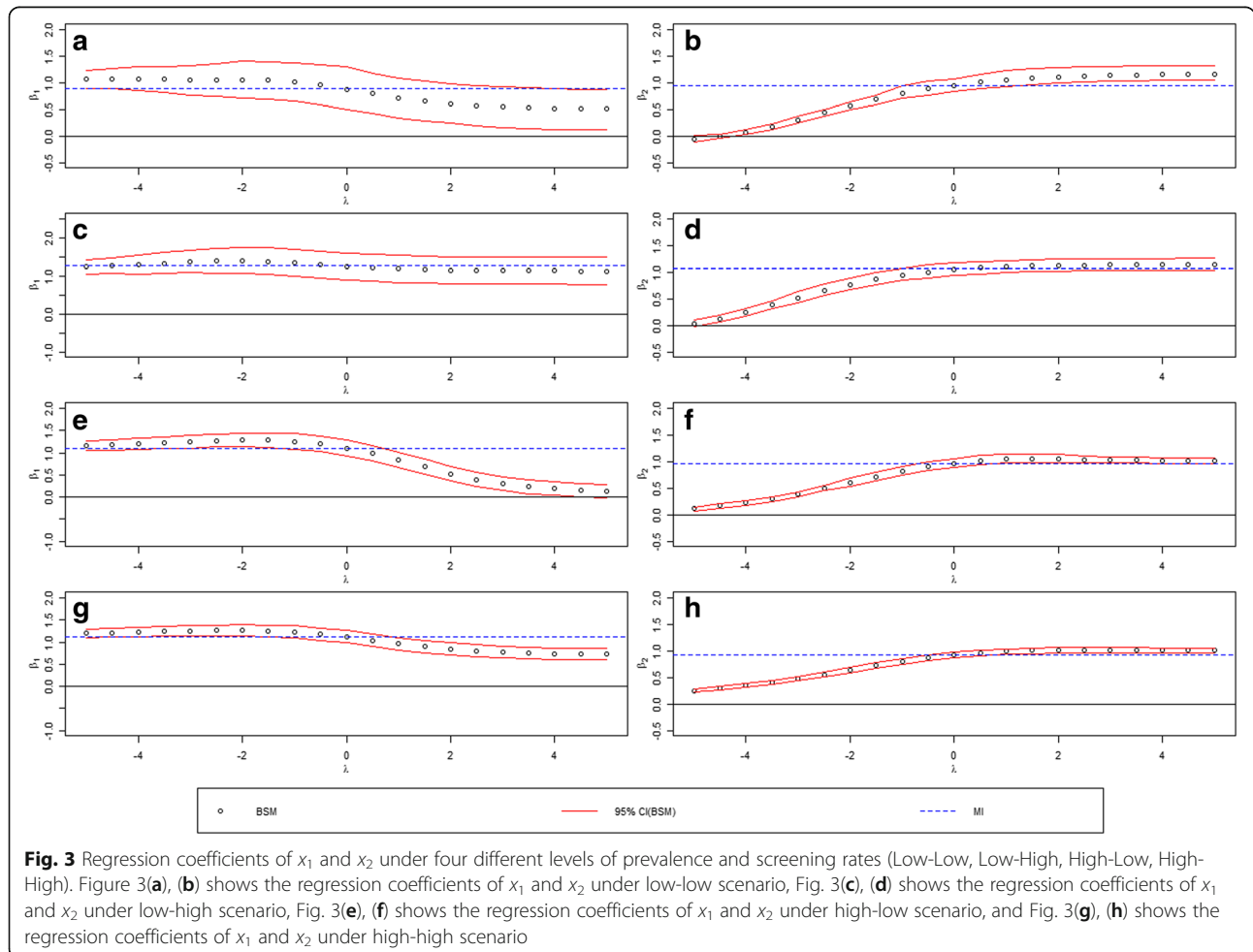
For the estimates of these risk factors (Fig. 1), the Bayesian selection model yields the same results as the complete-case (CC) analysis when $\lambda = 0$; this is not surprising because when the missingness depends on the covariates but not the outcome, the complete-case analysis is fully efficient for the regression [13]. When λ is sufficiently large (e.g., approaching to 4 in this example), the BSM methods yield results close to that of NNT. This is because when λ is large, it is sufficient to assume that those who did not receive chest radiography were negative for radiographic pneumonia. MI yields estimates close to CC for gender, fever at ED, severe hypoxia but smaller effect for age.

Figure 2 shows the overall prevalence of radiographic pneumonia in the study population decreases as sensitivity parameter λ increases. When λ is between 0 and 4, the estimates of the prevalence range from 0.056 to 0.032.

Simulation studies

The prevalence of radiographic pneumonia in the current study population is less than 6%, which is relatively low. In some other diseases such as sexually transmitted diseases and hypertension, the prevalence could be much higher. Logistic models are well-known to suffer from bias for rare events, and therefore the prevalence has an impact on the proposed BSM method [20]. On the other hand, missingness proportion plays an important role in the performance of missing data models, and therefore in our setting the performance of the BSM method could also be affected by the proportion of screening. In this section, we assess the performance of the BSM for different values of disease prevalence and proportions of subjects screened.

We generate two covariates, x_1 which is a binary variable from Bernoulli distribution (e.g., gender) and x_2 from



a normal distribution (e.g., age of high school students), and b_1, b_2 are the regression coefficients of x_1, x_2 in the main outcome model, respectively,

$$x_1 \sim \text{Bernoulli}(0.5)$$

$$x_2 \sim N(16, 2)$$

and we generate the disease status y based on a logistic regression model,

$$\text{logit}[\text{Pr}(y = 1)] = a + x_1 + x_2$$

and the response indicator R is also generated based on a logistic regression model,

$$\text{logit}[\text{Pr}(R = 1)] = b - x_1 + 0.5 * x_2 + 0.5 * y_1$$

We choose a to be -22.2 and -18.7 so the prevalence is around 2% (low prevalence) and 20% (high prevalence), and b to be -8.5 and -7.0 so that the response rates are 30% (low screening rate) and 60% (high screening rate). The combinations result in four simulation scenarios: (1) low prevalence and low screening rate; (2) low prevalence and high screening rate; (3) high prevalence and low screening rate; (4) high prevalence and high screening rate. We simulate 10,000 subjects from the study population and plot the regression estimates vs. the sensitivity parameter for each regression coefficient in Fig. 3, and the estimated prevalence vs. the sensitivity parameter in Fig. 4. To assess the performance of the BSM methods under the true sensitivity parameter ($\lambda = 0.5$), we replicate the process 200 times and evaluate

the methods by assessing the empirical bias, the root mean squared error, and the coverage probabilities of the 95% credible interval.

Table 1 shows the bias, RMSE and coverage probability of the 95% credible intervals of the proposed method along with NNT, CC and MI methods, when the sensitivity parameter is set at the true value (0.5). The true values of the regression coefficients of x_1 and x_2 are both 1, and the empirical bias is less than 1% for all but one coefficient, out of all simulation scenarios for the BSM method; Only the coefficient of x_2 shows more than 1% empirical bias under the scenario when both prevalence and the screening proportion are low. We see an improvement in the RMSEs (i.e., smaller RMSEs) with increase in either the disease prevalence or the screening proportion. As expected, the coefficient of the binary covariate x_1 has larger RMSEs than the coefficient of the continuous covariates x_2 . In general, the BSM method achieves good coverages for both regression coefficients. The two cases with coverage probabilities less than 90% are for the coefficient of x_2 when the screening probability is low. All other methods show large bias, increased RMSE and poor confidence coverage compared to BSM method.

Figure 3 shows the point estimates plots along with the 95% credible intervals of the two regression coefficients of x_1 and x_2 , under the four simulation scenarios. Figure 3(a), (c), (e) and (g) are for coefficients of x_1 under (1) low prevalence and low screening rate, (2) low prevalence and high screening rate, (3) high prevalence and low screening rate, (4) high prevalence and high

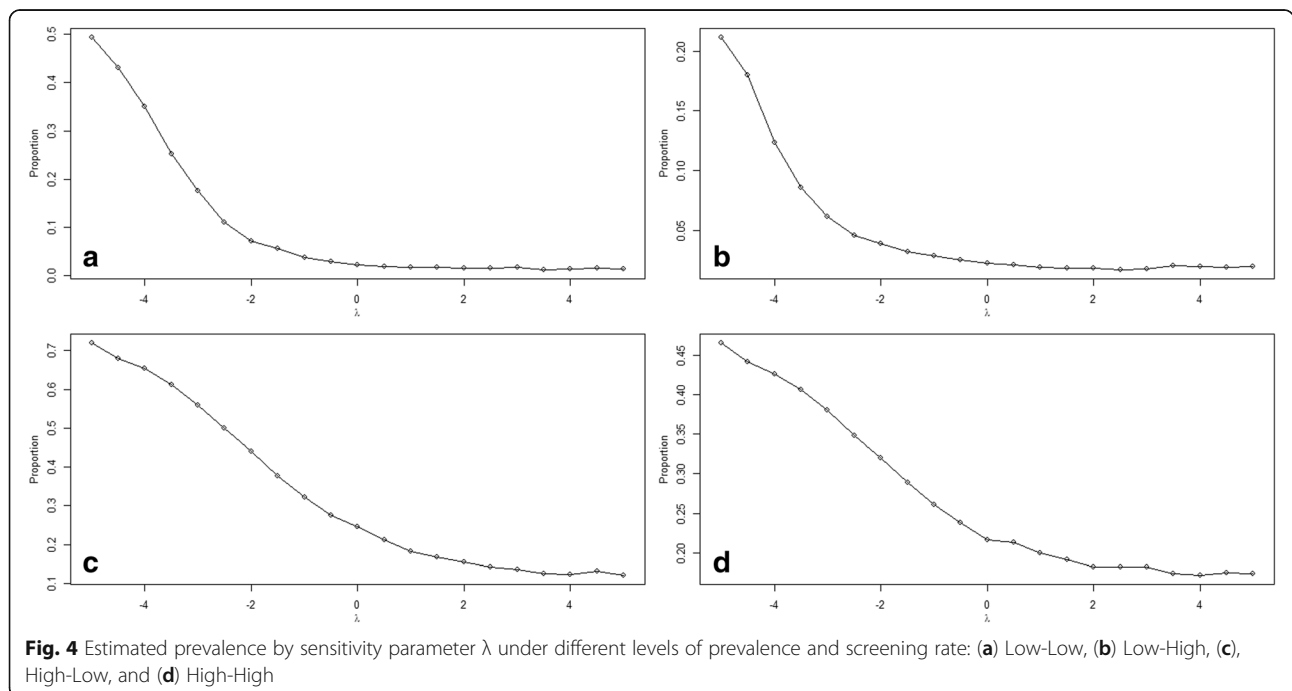


Table 1 Bias, RMSE, and Coverage Probability of 95% Credible Interval of the BSM, NNT, CC, and MI method, When Sensitivity Parameter is Set to True Value

			Low-Low	Low-High	High-Low	High-High
BSM	Bias*1000	b_1	3.19	4.14	1.91	2.88
		b_2	12.94	4.81	2.60	1.88
	RMSE*1000	b_1	192.51	163.04	94.14	66.59
		b_2	75.19	54.64	40.47	28.12
	Coverage Probability	b_1	96.0%	97.0%	97.0%	97.5%
		b_2	88.0%	94.5%	88.5%	91.0%
NNT	Bias*1000	b_1	-309.37	-101.24	-857.09	-331.60
		b_2	136.93	56.88	30.70	34.85
	RMSE*1000	b_1	362.4	190.85	860.28	338.08
		b_2	151.02	77.23	41.93	42.98
	Coverage Probability	b_1	63.0%	95.0%	0%	0%
		b_2	43.5%	86.0%	81.0%	73.5%
CC	Bias*1000	b_1	89.30	43.30	113.73	77.02
		b_2	-38.4	-24.67	-51.17	-37.12
	RMSE*1000	b_1	212.04	167.71	148.09	101.61
		b_2	81.55	60.70	64.04	45.83
	Coverage Probability	b_1	93.0%	96.0%	76.5%	83.5%
		b_2	89.5%	91.5%	68.5%	74.5%
MI	Bias*1000	b_1	97.01	43.23	115.96	79.41
		b_2	-43.32	-25.72	-51.34	-35.10
	RMSE*1000	b_1	221.17	168.02	150.53	103.85
		b_2	84.14	61.07	64.81	44.29
	Coverage Probability	b_1	92.5%	97.0%	76.0%	82.0%
		b_2	88.0%	92.5%	71.5%	76.0%

*denotes "multiplied by"

BSM Bayesian selection model, NNT Negative for not tested, CC Complete-case analysis, MI Multiple imputation

screening rate, respectively, while Fig. 3(b), (d), (f), (h) are for coefficient of x_2 under the corresponding four simulation scenarios. We see similar trends in the coefficient estimates with the change of the sensitivity parameter. However, there is substantial improvement in precision (tighter confidence bands) with the increase of prevalence or screening proportion. The estimated prevalence under different sensitivity parameters for the four simulation scenarios were plotted in Fig. 4.

Discussion

Under-screening results in missing disease status or misclassified disease status when assumptions are made for those who did not receive the screening test. The goal of our method was to demonstrate the use of the Bayesian selection model for missing outcome or misclassified outcome due to under-screening. Unlike other methods that rely on assumptions [8, 9] or validation data, [12] the BSM method relates the propensity of receiving screening to the disease status through a sensitivity

parameter. By varying the sensitivity parameter, the BSM method demonstrated how the prevalence and the association of risk factors change with the sensitivity parameter. We further used simulation studies to demonstrate the performance of BSM method under different levels of disease prevalence and screening proportion. Our simulation indicates that the BSM method performs well even under scenarios when both the prevalence and the screen proportion are low.

For illustration purpose, we applied the proposed BSM method to a pneumonia dataset. The results showed increased risk of pneumonia in girls, which is consistent with studies from Japan [21]. The results also indicated that having fever during ED or at triage, or having severe hypoxia, is positively associated with radiographic pneumonia. This is not surprising, as both fever and hypoxia are symptoms of pneumonia in kids [22]. A more rigorous analysis of the risk factors for radiographic pneumonia would need to examine more risk factors and possibly their interactions.

The Bayesian selection model is an important tool to consider for estimating the disease prevalence and in studying risk factors of the disease, when only a subset of the target population receive the test. For studying the association of the risk factors, i.e., the regression of outcome on risk factors, this method reduces to the complete-case analysis when the sensitivity parameter is set to zero, and approximates the NNT method when the sensitivity parameter approaches infinity. Unfortunately, there is no information available to estimate the sensitivity parameter without validation sample. The choice of the sensitivity parameter can be aided by gathering information relating the propensity of receiving the test to the actual disease status. The choice of covariates in the outcome model and the response indicator model can be aided by input from substantive experts regarding the hypothesized relationship of variables with the outcome and/or the response indicator. When validation data are available, it is possible to identify the parameters in the Bayesian selection model. In future work, we plan to study how to efficiently make use of the validation data.

Conclusions

In the current study, we developed a Bayesian selection model that jointly modeled the binary outcome and the response indicator for the case when the binary outcome may be missing or misclassified due to under-screening. The model for the response indicator relates the propensity of receiving screening to the disease status through a sensitivity parameter. The application of the model to a pneumonia data yielded results that were consistent with previous studies. The performance of the proposed method over other methods in the simulation studies demonstrated the promise of the proposed model for modeling missing or misclassified disease outcome due to under-screening.

Abbreviations

BSM: Bayesian selection model; CC: Complete-case analysis; CI: Confidence interval; COPD: Chronic obstructive pulmonary disease; ED: Emergency department; HCV: Hepatitis C virus; MAR: Missing at random; MCAR: Missing completely at random; MCMC: Markov chain Monte Carlo; MI: Multiple imputation; MNAR: Missing not at random; NNT: Negative for not tested; RMSE: Root mean squared error; STD: sexually transmitted disease

Acknowledgements

Not applicable

Funding

Not applicable

Availability of data and materials

The pneumonia dataset for this study was maintained by Cincinnati Children's Hospital. The investigators obtained approval of the Institutional Review Board to use the dataset. The dataset cannot be made publicly available because of protected health information (PHI) contained in the dataset. The simulation work was performed in R and is available from the corresponding author on request.

Authors' contributions

The study concept and design, as well as the interpretation of results were conducted by NZ, LA, TAF and MM. The programming, analysis of the manuscript were conducted by NZ and SC. All authors were involved in critical revision of the manuscript for important intellectual content. All authors gave final approval of the version to be published and agreed to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

Ethics approval and consent to participate

The institutional review board (IRB) at Cincinnati Children's Hospital approved the study and waived the need for informed consent.

Consent for publication

Not applicable

Competing interests

The authors declares that they have no competing interests.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Author details

¹Division of Biostatistics and Epidemiology, Cincinnati Children's Hospital Medical Center, 3333 Burnet Ave, MLC 5041, Cincinnati, OH 45229, USA.

²Department of Pediatrics, University of Cincinnati College of Medicine, Cincinnati, OH, USA. ³Division of Emergency Medicine, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, USA.

Received: 15 August 2017 Accepted: 27 November 2017

Published online: 12 December 2017

References

- Riley MR, Bass NM, Rosenthal P, et al. Underdiagnosis of pediatric obesity and underscreening for fatty liver disease and metabolic syndrome by pediatricians and pediatric subspecialists. *J Pediatr.* 2005;147(6):839–42.
- Cazzola M, Puxeddu E, Bettoncelli G, et al. The prevalence of asthma and COPD in Italy: a practice-based study. *Respir Med.* 2011;105(3):386–91.
- Hsieh Y-H, Rothman RE, Laeyendecker OB, et al. Evaluation of the Centers for Disease Control and Prevention recommendations for hepatitis C virus testing in an urban emergency department. *Clin Infect Dis.* 2016;62(9):1059–65.
- Shebl FM, El-Kamary SS, Shardell M, et al. Estimating incidence rates with misclassified disease status: a likelihood-based approach, with application to hepatitis C virus. *Int J Infect Dis.* 2012;16(7):e527–e31.
- Girardi E, Sabin CA, Antonella d'Arminio Monforte M. Late diagnosis of HIV infection: epidemiological features, consequences and strategies to encourage earlier testing. *JAIDS Journal of Acquired Immune Deficiency Syndromes.* 2007;46:53–8.
- Hansen ML, Gunn PW, Kaelber DC. Underdiagnosis of hypertension in children and adolescents. *JAMA.* 2007;298(8):874–9.
- Yamada K, Maeno T, Waza K, et al. Under-diagnosis of alcohol-related problems and depression in a family practice in Japan. *Asia Pacific family medicine.* 2008;7(1):1.
- Copeland KT, Checkoway H, McMichael AJ, et al. Bias due to misclassification in the estimation of relative risk. *Am J Epidemiol.* 1977; 105(5):488–95.
- Jurek AM, Greenland S, Maldonado G, et al. Proper interpretation of non-differential misclassification effects: expectations vs observations. *Int J Epidemiol.* 2005;34(3):680–7.
- Chyow P-H. Patterns of bias due to differential misclassification by case-control status in a case-control study. *Eur J Epidemiol.* 2007;22(1):7–17.
- Hausman JA, Abrevaya J, Scott-Morton FM. Misclassification of the dependent variable in a discrete-response setting. *J Econ.* 1998;87(2):239–69.
- Savoca E. Accounting for misclassification bias in binary outcome measures of illness: the Case of post-traumatic stress disorder in male veterans. *Social Methodol.* 2011;41(1):49–76.
- Little RJ, Zhang N. Subsample ignorable likelihood for regression analysis with missing data. *J R Stat Soc: Ser C: Appl Stat.* 2011;60(4):591–605.

14. Rubin DB. Multiple imputation for nonresponse in surveys: John Wiley & Sons; 2004.
15. Little RJ, Rubin DB. Statistical analysis with missing data: John Wiley & Sons; 2014.
16. Statistics NCFH. The international classification of diseases, 9th revision, clinical modification: procedures: tabular list and alphabetic index. US Department of Health and Human Services, Public Health Service, Health Care Financing Administration; 1980.
17. Florin TA, Carron H, Huang G, et al. Pneumonia in children presenting to the emergency department with an asthma exacerbation. *JAMA Pediatr.* 2016;
18. Schuh S, Lalani A, Allen U, et al. Evaluation of the utility of radiography in acute bronchiolitis. *J Pediatr.* 2007;150(4):429–33.
19. Lunn DJ, Thomas A, Best N, et al. WinBUGS-a Bayesian modelling framework: concepts, structure, and extensibility. *Stat Comput.* 2000;10(4):325–37.
20. King G, Zeng L. Logistic regression in rare events data. *Polit Anal.* 2001:137–63.
21. Eshima N, Tokumaru O, Hara S, et al. Age-specific sex-related differences in infections: a statistical analysis of national surveillance data in Japan. *PLoS One.* 2012;7(7):e42261.
22. Mahabee-Gittens EM, Grupp-Phelan J, Brody AS, et al. Identifying children with pneumonia in the emergency department. *Clin Pediatr.* 2005;44(5):427–35.

Submit your next manuscript to BioMed Central and we will help you at every step:

- We accept pre-submission inquiries
- Our selector tool helps you to find the most relevant journal
- We provide round the clock customer support
- Convenient online submission
- Thorough peer review
- Inclusion in PubMed and all major indexing services
- Maximum visibility for your research

Submit your manuscript at
www.biomedcentral.com/submit

