Published in final edited form as:

Am J Drug Alcohol Abuse. 2018 ; 44(2): 160–166. doi:10.1080/00952990.2017.1421212.

Comparing Methods of Misclassification Correction for Studies of Adolescent Alcohol Use

Melvin D. Livingston, Ph.D.¹, Brad Cannell, M.P.H., Ph.D.¹, Keith Muller, Ph.D.², and Kelli A. Komro, MPH, Ph.D.³

¹Department of Biostatistics and Epidemiology, School of Public Health, University of North Texas Health Science Center, Fort Worth, TX, USA

²Department of Health Outcomes & Policy and Institute for Child Health Policy, College of Medicine, University of Florida, Gainesville, FL, USA

³Department of Behavioral Sciences and Health Education, Rollins School of Public Health, Emory University, Atlanta, GA, USA

Abstract

Background—Despite concerns over measurement error, self-report continues to be the most common measure of adolescent alcohol use used by researchers. Objective measures of adolescent alcohol use continue to advance; however, they tend to be cost prohibitive for larger studies. By combining appropriate statistical techniques and validation subsamples, the benefits of objective alcohol measures can be made more accessible to a greater number of researchers.

Objectives—To compare three easily implemented methods to correct for measurement error when objective measures of alcohol use are available for a subsample of participants: regression calibration, multiple imputation for measurement error (MIME), and probabilistic sensitivity analysis (PSA), and provide guidance regarding the use of each method in scenarios likely to occur in practice.

Methods—This simulation experiment compared the performance of each method across different sample sizes, both differential and non-differential error, and differing levels of sensitivity and specificity of the exposure measure.

Results—Failure to adjust for measurement error led to substantial bias across all simulated scenarios ranging from a 35% to 208% change in the log-odds. For non-differential misclassification, regression calibration reduced this bias to between a 1% and 23% change in the log-odds regardless of sample size. At higher sample sizes, MIME produced approximately unbiased (between a 0% and 9% change in the log-odds) and relatively efficient corrections for

Correspondence: Melvin D. Livingston, Ph.D., 3500 Camp Bowie Blvd, Fort Worth, TX 76107, melvin.livingston@unthsc.edu.

Compliance with Ethical Standards and Financial Disclosures

Conflict of Interest

The authors report no conflicts of interest.

Ethical Approval

IRB approval is not applicable for this study.

Informed Consent

Informed Consent is not applicable for this study.

both non-differential and differential misclassification. PSA provided little utility for correcting misclassification due to the inefficiency of its estimates.

Conclusion—Concern over measurement error resulting from self-reported adolescent alcohol use persist in research. Where appropriate, methods involving validity subsamples provide an efficient avenue for addressing these concerns.

Keywords

Measurement error; misclassification; validity studies; substance use

Introduction

Misclassification in the measurement of adolescent alcohol use is a consistent source of bias. Although other forms of data collection are being advanced in the measurement of alcohol use(1–3), adolescent alcohol research remains primarily driven by self-report surveys(4). The impact of self-report on the measurement of adolescent alcohol use varies with type and structure of the question being asked(5). For example, self-report measures of recent alcohol use have been found to have variable validity when compared to objective measures of alcohol use (69% to 93% consistency)(6). Shillington et al found high test-retest reliability of any alcohol use among adolescents (94% consistency after two years). However, Shillington et al also found that, after two years, only 48% of adolescents reported consistent ages of alcohol use onset allowing for a single year of discrepancy improving to 80% when allowing for two years of discrepancy(7). Additionally, consistency of reporting age at first alcohol use has been shown to be associated with socioeconomic status and recent substance use(8).

Self-report measures are the most practical means of obtaining information on the frequency, amount, and age of adolescent alcohol use in non-clinical research settings(4). Ideally, researchers would be able to use the best available measurement tool on all participants. For example, instead of relying self-reported alcohol in the past 30 days, participants could be monitored in real time with ecological momentary assessments or wearable technology. Likewise, instead of relying on items with long recall periods such as age of alcohol use initiation, participants could be assessed regularly from a young age prospectively. While using such measures would certainly decrease any bias resulting from measurement error, they are often not feasible for all participants in a study either due to logistical complications or cost(5). However, in circumstances that allow for a subset of study participants' alcohol use to be measured with little or no error, statistical procedures exist that incorporate information from these validation sub-samples to reduce bias in estimated effects of alcohol use measures on associated outcomes.

Objective measures of alcohol use continue to improve. However, examples of their use in behavioral studies of adolescents remains rare. As one example, Dougherty et al (2015) have demonstrated how transdermal alcohol monitors, previously used in the criminal justice system, might be used for research purposes(2). Until validated biomarkers or transdermal alcohol monitors are inexpensive enough to replace survey responses for all study participants in behavioral and intervention research, it is still logistically and financially

optimal to use these techniques for a limited sub-sample of participants. As a result, it is important for researchers to understand how to efficiently incorporate validity sub-study data into the design and analysis of adolescent alcohol use outcomes.

Existing methods for correcting bias due to misclassification when a validity sub-sample is present vary greatly in their sophistication and underlying assumptions, including: probabilistic sensitivity analyses(9), regression calibration(10–13), maximum likelihood(14–16), a variety of nonparametric methods(17,18), Bayesian methods(19–21), and multiple imputation methods(22). Despite their potential utility, it is uncommon to see these methods implemented in practice – perhaps due to the complexity of their implementation. However, the following methods, which are the topic of the current study, are easily implemented via existing software: regression calibration, multiple imputation for measurement error (MIME), and probabilistic sensitivity analysis (PSA).

In order for researchers to make best use of validation sub-study data to correct for misclassification, they will need to understand both how available methods correct for misclassification, their utility in different scenarios that may be experienced in practice, and have access to simple examples demonstrating their use. Additionally, it is of importance that researchers are aware of the full extent of potential bias when misclassification is ignored. This study provides a summary of the methodology and example code for three procedures for misclassification correction that are easy to implement: regression calibration, MIME, and PSA. We also present the results of a simulation study for binary misclassification comparing each of these methods to a naïve approach with no correction for misclassification, and make recommendations about when they are best used in practice.

Methods

Below we briefly describe three existing analytical methods that can be used along with validation subsamples. Throughout the descriptions we refer to the potentially misclassified full-sample measure as the exposure subject to error (ESE) and the gold standard measure found only in the validity subsample as the validation measure. The three methods presented can be easily implemented using existing tools in SAS, and we provide citations and links to existing example code for the use of the reader. Additionally, example SAS code can be found in the appendix demonstrating the use of each method.

For the sake of simplicity in our simulations, we focus on the use of these methods for correcting binary misclassification. However, some methods can be easily modified to account for measurement error in variables with other distributions. Modifications to the example code are briefly discussed.

Regression calibration

Regression calibration works by fitting a series of regression models. First, the primary analysis model is fit with the full sample estimating the association between the outcome of interest and the ESE. To correct for misclassification, a linear regression is then run estimating the association between the validation measure and the ESE. From this second regression, correction factors are calculated that are then applied to the results from the

primary analysis. An advantage to the regression calibration method is that while it requires a linear relationship between the ESE and the validation measure, it does not require normality of the residuals in the error model making it useful for correcting measurement error in both binary and continuous exposures. The primary disadvantage for regression calibration is that it is not compatible data that has been differentially misclassified across the outcome of interest.

This method can be easily implemented in SAS by hand, and example code is given in the appendix.

Multiple imputation for measurement error

MIME reframes the issue of measurement error as a missing data problem. The ESE is present for all participants, while the validation measure is only present for a random sample of participants. As long as that validation sample is truly a random sample of the study participants, then the missing validation measures are missing completely at random (MCAR). For data that is MCAR multiple imputation can be used to estimate what the validation measure would have been across all participants conditional on the ESE and outcome giving an efficient and approximately unbiased estimate of the true association between exposure and the outcome of interest.

MIME is easily implemented using the multiple imputation tools available in SAS: PROC MI and PROC MIANALYZE. The logic of any multiple imputation analysis follows three steps. First multiple complete data sets are imputed based on the variables supplied to PROC MI. The user then runs the desired analysis in each of the generated imputation data sets saving the results to combine in the next step. Finally, PROC MIANALYZE combines the estimates from each imputation data set. To minimize bias and maximize efficiency, it has been suggested that at least 40 imputations are generated when using MIME(22).

The primary advantage to the MIME approach is its flexibility. By modifying the variables supplied to PROC MI to include an interaction term between the outcome of interest and the ESE, differential misclassification is easily accommodated. Additionally, PROC MI can explicitly impute continuous, binary, and non-binary categorical data allowing users to implement MIME across these variable types.

Probabilistic sensitivity analysis

The probabilistic sensitivity analysis method attempts to recreate the data that would have existed had the exposure variable not been misclassified. It does this by reclassifying the existing ESE based on a set of external distributions for the sensitivity and specificity of the misclassified measure. For non-differential misclassification, two external distributions are chosen: one for the specificity and one for the sensitivity of the measure. For differential misclassification, four external distributions are specified: the sensitivity and specificity of cases and the sensitivity and specificity of controls. A single iteration of the simulation draws from these distributions a set of sensitivity and specificity values. These values are then used to calculate the positive predictive value (PPV) and negative predictive value (NPV) of exposure classification. These values are then applied to the corresponding individual records. Next, a random number is generated from a uniform(0,1) distribution for

each record. If this number is larger than the records probability of being correctly classified then the record is reclassified. Finally a logistic regression is run on the newly classified data, and a summary log odds ratio calculated. In order to account for random error the standard error for the conventional log odds ratio is calculated; then a value is sampled from a standard normal distribution. The product of this standard normal deviate and the conventional standard error is subtracted from the reclassified log odds ratio. This process is then repeated many times resulting in a distribution of odds ratios adjusted for both random and systematic error.

This method can be implemented using the SAS macro %SENSMAC written by Fox, Lash and Greenland and available at <https://sites.google.com/site/biasanalysis/sensmac>.⁽⁹⁾ %SENSMAC, provides for the misclassification rates to be specified under a uniform, triangular, or trapezoidal distribution. Fox and associates' discussion of the uses of this method do not explicitly include situations where a validation sub-study is present. In our implementation using validation sub-samples, the triangular distribution is specified, with the peak determined by the misclassification rates calculated within the validation sample, and the upper and lower bounds set to 1.96 standard errors above and below the peak. We allow for separate sensitivity and specificity distributions by outcome status in all analyses using probabilistic sensitivity analyses to allow for differences in the sensitivity and specificity due to random error.

Simulation Design

The purpose of this simulation study is to compare the relative performances of the methods described with respect to bias, random error, and interval coverage. To achieve this goal, 2000 simulation datasets were generated for each of four different misclassification scenarios at three different sample sizes. Two non-differential misclassifications scenarios were tested: a high sensitivity and specificity scenario where both the sensitivity and specificity were set to 0.9, and a low sensitivity and specificity scenario where the sensitivity and specificity were set to 0.6. Two differential misclassifications scenarios were also tested: a high sensitivity and specificity in the cases with a high sensitivity and low specificity in the controls scenario, and low sensitivity and specificity in the cases with a low sensitivity and high specificity in the controls scenario. These values were selected to provide a range of misclassification values that might be seen in practice. High sensitivity and specificity would be expected for questions about recent alcohol use, while low sensitivity and specificity has been observed for long-term recall questions such as age at first alcohol use. To evaluate these methods across the varying study sizes seen in adolescent alcohol research, we tested each method at sample sizes of 200, 1000, and 2000^(23–26). We restricted the number of misclassification scenarios due to the large amount of computing time required for each scenario. The scenarios that were presented were chosen to demonstrate the performance of each method in a practical best and worst case scenario with regards to misclassification rates.

Each simulation iteration dataset consisted of the following variables: X, Y, V, and ESE; where X was the true exposure measured without error simulated from a distribution Bernoulli(p_1). Y was the outcome measured without error. Y was constructed so that the

relationship between X and Y produces a change in the log-odds of 0.69 (odds ratio of 2.0) under logistic regression. In order to simulate the presence of a validation sub-study, the variable V was constructed by randomly sampling 30% of the records in a simulation dataset; those sampled had V set to the corresponding value for X, while those not sampled had V set to missing. ESE represented the misclassified exposure. ESE was constructed from X by purposefully misclassifying X according to specified sensitivity and specificity values (an example of this can be seen in the code provided in the appendix).

For each of the 2000 simulation datasets in the scenarios described, we estimated the bias, standard error, and interval coverage of each method. The bias for each method was calculated by taking the difference between the log odds estimate from the correction method and the true log odds. The standard errors of the naïve approach, regression calibration and MIM are obtained directly from the output. The standard error for PSA was calculated by assuming that the resulting interval can be viewed as a confidence interval based on a normal distribution and back calculating the standard error from the upper and lower interval estimates.

Results

Results for our simulation analyses can be found in Table 1. For ease of interpretation we expressed bias in two separate ways. First, we calculated the percent change in the estimated log odds for each correction method to show the average relative deviation from the true estimate. Second, we estimated what the average observed odds ratio would have been for each correction method in each of the simulated scenarios. By comparing the observed odds ratio to the true odds ratio of 2.0, the reader can easily see the impact of ignoring misclassification error as well as the effect of each correction method. Interval coverage is expressed as the percentage of simulated 95% confidence intervals that contain the true estimate. For a 95% confidence interval, an ideal estimator would contain the true point estimate approximately 95% of the time. Deviations from 95% coverage indicate some flaw in the correction method: either it leads to biased estimates or it produces confidence intervals that are either overly narrow or wide. Results for each method are discussed in detail below.

Naïve Analysis (Uncorrected for Misclassification)

Ignoring misclassification lead to substantial bias and poor confidence interval coverage in all scenarios tested. Even in our best case scenario where misclassification was nondifferential and sensitivity and specificity were both high, we observed a 35% reduction in the change in log odds resulting in an underestimated odds ratio of 1.57 (compared to the true odds ratio of 2.00). Under differential misclassification, substantial bias was observed either greatly overestimating the true odds ratio (3.94 vs 2.00) or reversing the directionality of the estimate (0.47 vs 2.00). While the bias in the estimated odds ratios remained fairly consistent across sample sizes, the interval coverage became markedly worse as the increased sample size allowed for a more precise estimate of the incorrect odds ratio.

Regression Calibration

Under non-differential misclassification, regression calibration produced less biased estimates than would be expected when ignoring misclassification that generally improved with increasing sample size. Under the best case scenario of high sensitivity and specificity, regression calibration produced approximately unbiased odds ratios across all sample sizes (1.95 to 1.97). Similarly, confidence interval coverage was near the nominal level of 0.95 across all sample sizes. More biased odds ratios were observed with more severe misclassification rates with an observed odds ratio of 1.70 when the sample size was at 200, improving to 1.88 with a sample size of 1000 and 1.93 with a sample size of 2000. Corresponding improvements were also seen in interval coverage. Unsurprisingly, regression calibration performed extremely poorly in all of the differential misclassification scenarios producing odds ratios ranging from 0.19 to over 200.

MIME

MIME was unable to produce any estimates when the sample size was 200 due to a lack of convergence in the imputation model. However, when sample sizes were high enough to enable adequate model convergence, MIME consistently produced odds ratio estimates close to the true odds ratio of 2.00. Observed odds ratio estimates across both differential and non-differential misclassification ranged between 1.88 and 2.00. Similarly, all estimated scenarios produced interval coverage close to the nominal level of 95%.

PSA

Similar to MIME, PSA was unable to produce estimates when the sample size was 200 due to the sparseness of the crosstab cells used to estimate the sensitivity and specificity of the sample. In terms of bias, the performance of the PSA method varied greatly with misclassification severity and sample size. PSA performed worst when both sensitivity and specificity were low producing observed odds ratio of 1.48 with a sample size of 1000 and 1.62 with a sample size of 2000. PSA performed best with high sensitivity and specificity, producing approximately unbiased odds ratios (1.99) at both sample sizes. While bias varied across the simulation scenarios, interval coverage was consistently higher than 95% with all scenarios having 100% interval coverage. Such a high level of consistent interval coverage indicates overly wide simulation-based intervals making any sort of statistical inference difficult due to the decreased statistical power.

Discussion

Despite concerns of misclassification, self-report measures of adolescent alcohol use continues to be a common approach for measuring adolescent alcohol use. The results presented for this naïve approach demonstrate the severity of the problem with relying on pure self-report data. In the best case scenario of non-differential misclassification with high sensitivity and specificity, ignoring misclassification lead to a reduction in the odds ratio from 2.0 to 1.6. In all other scenarios, the bias was severe enough to completely change any substantive interpretation. Interpretation of results that rely purely on self-report of adolescent alcohol use alone should be done cautiously.

While the discussed methods for misclassification correction were able to minimize the bias due to misclassification under some circumstances, they are not without limitations that may impact their use. Of the three methods discussed, only regression calibration was able to provide unbiased estimates with sample sizes of 200, and only for circumstances where the misclassification was nondifferential. Given the resulting bias observed when misclassification was ignored, we urge researchers to carefully consider and discuss the potential impact of misclassification particularly with smaller studies where options for correction are limited. While the methods discussed are of limited use with smaller studies, intervention studies of adolescent alcohol use often include sample sizes in excess of 1000(23–26). Subsequent secondary data analysis of such trials exploring etiological questions is also quite common, and plays an important role in the adolescent alcohol literature(27,28). We encourage researchers to incorporate validation subsamples into these larger trials to enable both the estimation of unbiased intervention effects as well as to improve our understanding of risk and protective factors associated with alcohol use that comes from secondary data analysis of these trials.

At sufficient samples sizes, MIME provided a flexible alternative for correction of misclassification error. MIME produced approximately unbiased odds ratio estimates and acceptable confidence interval coverage regardless of whether the ESE was misclassified differentially or non-differentially with respect to the outcome. A unique advantage of MIME compared to the other methods discussed is that frames measurement error in the broader context of missing data. For researchers already familiar with the use of multiple imputation for dealing with missing data, this allows MIME to be easily used without a need for additional software or “by hand” coding. Additionally, by nesting measurement error in the context of missing data, MIME can also be extended to dealing with measurement error in the presence of unplanned missing data common to adolescent alcohol research by extending the imputation model to include other variables containing missing data.

PSA performed unexpectedly poorly in comparison to regression calibration and MIME. While it did consistently produce less biased estimates than the naïve approach, the width of PSA’s simulation based interval was much larger than the confidence intervals of the other methods. The intervals produced in our simulation study were similar in width to the example shown by Fox et al (2005). One notable difference between the example shown in Fox et al. and our simulations is the magnitude of the true effect estimate. Fox and associates’ true effect estimate was an odds ratio of approximately 3.0, while our true effect estimate was an odds ratio of 2.0. The only situation in which PSA provided comparable estimates to the other methods was when the misclassification rates were already very low. While PSA may be useful purely as a sensitivity analysis, we cannot recommend its use as a method to correct for misclassification in the current implementation.

In addition to sample size requirements, the primary limitation of these procedures is the necessity of unbiased estimates of the relationship between the ESE and the validation measure. For regression calibration and MIME, this requires a validation sub-sample to compare the ESE to the validation measure. For PSA, the sensitivity and specificity of the misclassified measure can either be calculated directly from a validation sub-sample or assumed based on prior literature. When no gold standard exists, each of these methods may

fail to reduce the bias in the estimation of the exposure-outcome relationship. If attempts are made to use these methods to correct for misclassification when the validation measure is itself misclassified, then it is possible to increase the bias in the estimation of the exposure-outcome relationship. If the validation measure cannot be assumed to be measured without error, then results from the misclassification correction procedures should be considered as results from sensitivity analyses and not definitive in and of themselves.

In conclusion, validity sub-samples can be an efficient way to decrease the bias resulting from using commonly misclassified alcohol use measures. The importance of these methods to adolescent alcohol research will only continue to grow as objective alcohol measures improve. With sufficient sample sizes, MIME is a particularly effective way to account for exposure misclassification. Future research should focus on improving accessible methods for misclassification correction for studies with small sample sizes.

Acknowledgments

Funding

The authors gratefully acknowledge funding support from the National Institute on Alcohol Abuse and Alcoholism (R01AA016549).

References

1. Dahl H, Voltaire Carlsson A, Hillgren K, Helander A. Urinary Ethyl Glucuronide and Ethyl Sulfate Testing for Detection of Recent Drinking in an Outpatient Treatment Program for Alcohol and Drug Dependence. *Alcohol Alcohol*. 2011 May 1; 46(3):278–82. [PubMed: 21339184]
2. Dougherty DM, Hill-Kapturczak N, Liang Y, Karns TE, Lake SL, Cates SE, et al. The Potential Clinical Utility of Transdermal Alcohol Monitoring Data to Estimate the Number of Alcoholic Drinks Consumed. *Addict Disord Their Treat*. 2015 Sep; 14(3):124–30. [PubMed: 26500459]
3. Kip MJ, Spies CD, Neumann T, Nachbar Y, Alling C, Aradottir S, et al. The Usefulness of Direct Ethanol Metabolites in Assessing Alcohol Intake in Nonintoxicated Male Patients in an Emergency Room Setting. *Alcohol Clin Exp Res*. 2008 Jul; 32(7):1284–91. [PubMed: 18540912]
4. Litten RZ, Bradley AM, Moss HB. Alcohol Biomarkers in Applied Settings: Recent Advances and Future Research Opportunities: ALCOHOL BIOMARKERS IN APPLIED SETTINGS. *Alcohol Clin Exp Res*. 2010 Apr 5; 34(6):955–67. [PubMed: 20374219]
5. Del Boca FK, Darkes J. The validity of self-reports of alcohol consumption: state of the science and challenges for research. *Addiction*. 2003 Dec; 98(Suppl 2):1–12.
6. Midanik L. The Validity of Self-Reported Alcohol Consumption and Alcohol Problems: A Literature Review. *Br J Addict*. 1982; (77):357–82. [PubMed: 6762224]
7. Shillington AM, Clapp JD. Self-report stability of adolescent substance use: are there differences for gender, ethnicity and age? *Drug Alcohol Depend*. 2000 Jul 1; 60(1):19–27. [PubMed: 10821986]
8. Livingston MD, Xu X, Komro KA. Predictors of Recall Error in Self-Report of Age at Alcohol Use Onset. *J Stud Alcohol Drugs*. 2016 Sep; 77(5):811–8. [PubMed: 27588540]
9. Fox MP. A method to automate probabilistic sensitivity analyses of misclassified binary variables. *Int J Epidemiol*. 2005 Jul 28; 34(6):1370–6. [PubMed: 16172102]
10. Spiegelman D, McDermott A, Rosner B. Regression calibration method for correcting measurement-error bias in nutritional epidemiology. *Am J Clin Nutr*. 1997 Apr; 65(4 Suppl): 1179S–1186S. [PubMed: 9094918]
11. Spiegelman D, Carroll RJ, Kipnis V. Efficient regression calibration for logistic regression in main study/internal validation study designs with an imperfect reference instrument. *Stat Med*. 2001 Jan 15; 20(1):139–60. [PubMed: 11135353]

12. Spiegelman D, Rosner B, Logan R. Estimation and Inference for Logistic Regression with Covariate Misclassification and Measurement Error in Main Study/Validation Study Designs. *J Am Stat Assoc.* 2000 Mar;95(449):51.
13. Thurston SW, Williams PL, Hauser R, Hu H, Hernandez-Avila M, Spiegelman D. A comparison of regression calibration approaches for designs with internal validation data. *J Stat Plan Inference.* 2005 Apr; 131(1):175–90.
14. Breslow NE, Holubkov R. Weighted likelihood, pseudo-likelihood and maximum likelihood methods for logistic regression analysis of two-stage data. *Stat Med.* 1997 Feb 15; 16(1–3):103–16. [PubMed: 9004386]
15. Satten GA, Kupper LL. Inferences About Exposure-Disease Associations Using Probability-of-Exposure Information. *J Am Stat Assoc.* 1993 Mar;88(421):200.
16. Spiegelman D, Casella M. Fully parametric and semi-parametric regression models for common events with covariate measurement error in main study/validation study designs. *Biometrics.* 1997 Jun; 53(2):395–409. [PubMed: 9192443]
17. Cheng J, Small DS, Tan Z, Ten Have TR. Efficient nonparametric estimation of causal effects in randomized trials with noncompliance. *Biometrika.* 2009 Jan 24; 96(1):19–36.
18. Pepe MS, Fleming TR. A Nonparametric Method for Dealing With Mismeasured Covariate Data. *J Am Stat Assoc.* 1991 Mar;86(413):108.
19. Dellaportas P, Stephens DA. Bayesian Analysis of Errors-in-Variables Regression Models. *Biometrics.* 1995 Sep;51(3):1085.
20. Prescott GJ, Garthwaite PH. A Bayesian approach to prospective binary outcome studies with misclassification in a binary risk factor. *Stat Med.* 2005 Nov 30; 24(22):3463–77. [PubMed: 16237661]
21. Richardson S, Gilks WR. A Bayesian approach to measurement error problems in epidemiology using conditional independence models. *Am J Epidemiol.* 1993 Sep 15; 138(6):430–42. [PubMed: 8213748]
22. Cole SR, Chu H, Greenland S. Multiple-imputation for measurement-error correction. *Int J Epidemiol.* 2006 Aug; 35(4):1074–81. [PubMed: 16709616]
23. Komro KA, Livingston MD, Wagenaar AC, Kominsky TK, Pettigrew DW, Garrett BA. Multilevel Prevention Trial of Alcohol Use Among American Indian and White High School Students in the Cherokee Nation. *Am J Public Health.* 2017 Jan;19:e1–7.
24. Skara S, Sussman S. A review of 25 long-term adolescent tobacco and other drug use prevention program evaluations. *Prev Med.* 2003 Nov; 37(5):451–74. [PubMed: 14572430]
25. Lemstra M, Bennett N, Nannapaneni U, Neudorf C, Warren L, Kershaw T, et al. A systematic review of school-based marijuana and alcohol prevention programs targeting adolescents aged 10–15. *Addict Res Theory.* 2010 Jan 1; 18(1):84–96.
26. Petrie J, Bunn F, Byrne G. Parenting programmes for preventing tobacco, alcohol or drugs misuse in children <18: a systematic review. *Health Educ Res.* 2007 Apr; 22(2):177–91. [PubMed: 16857779]
27. Tobler AL, Livingston MD, Komro KA. Racial/ethnic differences in the etiology of alcohol use among urban adolescents. *J Stud Alcohol Drugs.* 2011 Sep; 72(5):799–810. [PubMed: 21906507]
28. Komro KA, Livingston MD, Garrett BA, Boyd ML. Similarities in the Etiology of Alcohol Use Among Native American and Non-Native Young Women. *J Stud Alcohol Drugs.* 2016 Sep; 77(5): 782–91. [PubMed: 27588537]

Table 1

Simulation results comparing methods of misclassification correction

Cases			Naïve (Uncorrected)				Regression Calibration				MIME				Probabilistic Sensitivity Analysis			
Sensitivity, Specificity	Sensitivity, Specificity	Control	Bias	Standard Error	MSE	Interval Coverage	Bias	Standard Error	MSE	Interval Coverage	Bias	Standard Error	MSE	Interval Coverage	Bias	Standard Error	MSE	Interval Coverage
N=200																		
Non-Differential	0.6, 0.6	0.6, 0.6	-0.63	0.28	0.48	0.45	-0.16	2.68	11.00	0.99	-	-	-	-	-	-	-	-
	0.9, 0.9	0.9, 0.9	-0.24	0.33	0.18	0.89	-0.02	0.51	0.35	0.96	-	-	-	-	-	-	-	-
Differential	0.6, 0.6	0.6, 0.9	0.68	0.34	0.58	0.46	4.67	2.64	30.72	0.67	-	-	-	-	-	-	-	-
	0.9, 0.9	0.9, 0.6	-1.44	0.30	2.16	0.00	-2.36	0.74	6.16	0.08	-	-	-	-	-	-	-	-
N=1000																		
Non-Differential	0.6, 0.6	0.6, 0.6	-0.61	0.13	0.39	0.00	-0.06	1.06	1.81	0.98	0.00	0.31	0.13	0.96	-0.30	1.81	3.41	1.00
	0.9, 0.9	0.9, 0.9	-0.24	0.15	0.08	0.64	-0.01	0.23	0.07	0.95	-0.04	0.26	0.09	0.97	0.00	0.43	0.18	1.00
Differential	0.6, 0.6	0.6, 0.9	0.68	0.15	0.57	0.00	4.67	1.15	23.12	0.00	0.01	0.29	0.12	0.95	0.12	1.40	1.97	1.00
	0.9, 0.9	0.9, 0.6	-1.43	0.13	2.05	0.00	-2.37	0.34	5.73	0.00	-0.06	0.28	0.11	0.96	-0.06	1.15	1.33	1.00
N=2000																		
Non-Differential	0.6, 0.6	0.6, 0.6	-0.60	0.09	0.37	0.00	-0.03	0.73	0.84	0.98	0.00	0.21	0.06	0.95	-0.21	1.46	2.17	1.00
	0.9, 0.9	0.9, 0.9	-0.24	0.10	0.07	0.38	-0.01	0.16	0.03	0.95	-0.01	0.18	0.04	0.96	0.00	0.27	0.08	1.00
Differential	0.6, 0.6	0.6, 0.9	0.68	0.11	0.48	0.00	4.65	0.81	22.30	0.00	0.00	0.20	0.06	0.95	0.04	0.98	0.97	1.00
	0.9, 0.9	0.9, 0.6	-1.43	0.09	2.04	0.00	-2.35	0.24	5.58	0.00	-0.01	0.19	0.05	0.96	-0.01	0.77	0.59	1.00