Correcting outcome misclassification in epidemiological studies: a simulation study

USTUR-0742-26A



Stacey L. McComish[1], Xirui Liu[1], Joey Y. Zhou[2], Sergey Y. Tolmachev[1]

[1] *United States Transuranium and Uranium Registries, College of Pharmacy and Pharmaceutical Sciences, Washington State University, Richland, WA, USA*

[2] *Office of Health Studies and Former Worker Programs, United States Department of Energy, Washington, D.C., USA*

Abstract

Death certificate errors are a common source of uncertainty in radiation epidemiological studies. Approximately 25% of death certificate causes of death from US Transuranium and Uranium Registries donors do not match their corresponding autopsy reports. Recent research has indicated that the impact of these errors on dose-response modeling may not be trivial. This simulation study applied a correction method published by Rogan and Gladen in 1978 to examine how misclassification and correction affected the significance of simulated dose–response associations. Preset underlying odds ratios ranging from 1 to 3.5 were used to generate 10,000 population, misclassified, and corrected datasets. Misclassified datasets were corrected by fitting each dataset using a logistic regression model, and determining the probability of disease as a function of dose. These probabilities were then corrected and used to generate corrected datasets. Results indicated that while the correction method successfully recovered the underlying odds ratio associations, the corrected odds ratio distributions became wider than misclassified odds ratio distributions, leading to inflated false positive and false-negative rates. For example, the false-positive rate can be simulated by generating datasets with an underlying odds ratio of 1, and determining the proportion that have significant p-values. When this was done, the false positive rate was 5% among both population and misclassified datasets, but increased to 21% as a consequence of the correction method. Similarly, false negative rates were simulated using underlying odds ratios that were greater than 1, and determining the proportion of non-significant p-values. Misclassification consistently led to false negative rates that were too high. The correction method attenuated this, resulting in false negative rates that were closer to the true population values, though they could still be too large. False negative rates decreased as the underlying odds ratio increased. This trend held true for the corrected false negative rates, which decreased as the underlying odds ratio increased.

Background

- 25% of USTUR death certificates do not match autopsy reports.
- Heuristic says non-differential misclassification shifts odds ratios toward the null.
- However, sometimes odds ratios move away from the null enough to change study conclusions from non-significant to significant.
- Rogan and Gladen described a method for estimating the true prevalence of disease by correcting the observed frequency of disease.

Simulation Methods

Starting point
- Preset underlying odds ratio (OR) between 1.0 and 3.5

TRUE odds ratio
What epidemiological studies are trying to determine

Generate 1 dataset
- 5,000 individuals
- External radiation doses (0-1 Gy)
- 20% baseline cancer rate

Repeat 10,000 times

POPULATION datasets
Represent what people actually died from

Misclassify each dataset
- 5% over-misclassification
- 10% under-misclassification

MISCLASSIFIED datasets
Represent what death certificates say people died from

Correct each dataset
- Apply Rogan-Gladen estimator to continuous dose data

CORRECTED datasets

Logistic fit
Fit each misclassified dataset using a logistic regression

Determine misclassified outcome probabilities
Use fit to calculate outcome probabilities for all 5,000 doses in each dataset

Correct outcome probabilities
Correct each misclassified probability using the Rogan-Gladen estimator

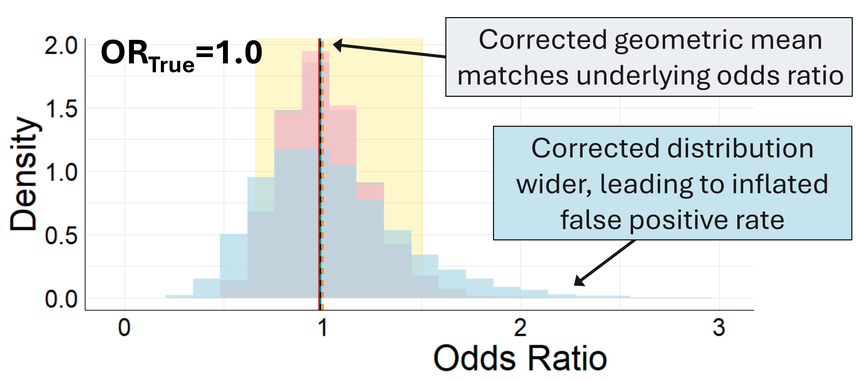
$$P_{Corr} = \frac{P_{Mis} - over}{1 - over - under}$$

Generate Corrected Outcomes
Use the binomial distribution to generate a corrected outcome for each corrected probability

Logistic fit to corrected datasets
Use fits to obtain β_1 regression coefficients and their 95% confidence intervals, where β_1=ln(OR)

Corrected False Positive Rate Too High

- False positives were assessed by starting with an underlying odds ratio of 1.0 (e.g. no association between dose and disease).
- On average, misclassification did not change the odds ratio.
- Misclassification also did not change the false positive rate associated with 95% confidence intervals (5%).
- The correction method increased the false positive rate from approximately 5% to 21%.

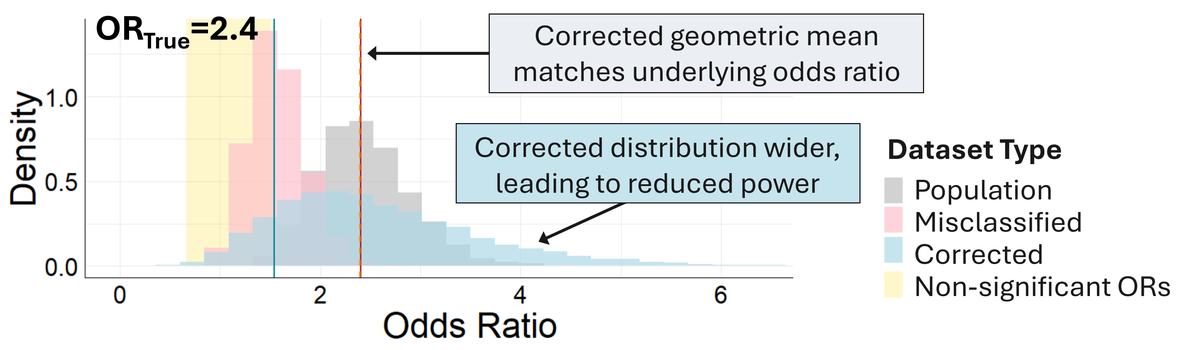


Dataset type	FPR†
Population	5.2%
Misclassified	4.8%
Corrected	21.4%

†FPR=False Positive Rate

Tradeoff: Correction Reduced Bias but Increased Standard Deviation for True OR > 1

- Misclassification shifts odds ratios toward the null, on average, decreasing the statistical power (true positive rate).
- Correction reduces this bias, but increases the standard deviation.
- In these simulations, the wider corrected distribution was associated with higher power than the population distribution for $1 < OR_{True} < 1.5$ and lower power for $OR_{True} > 1.5$.



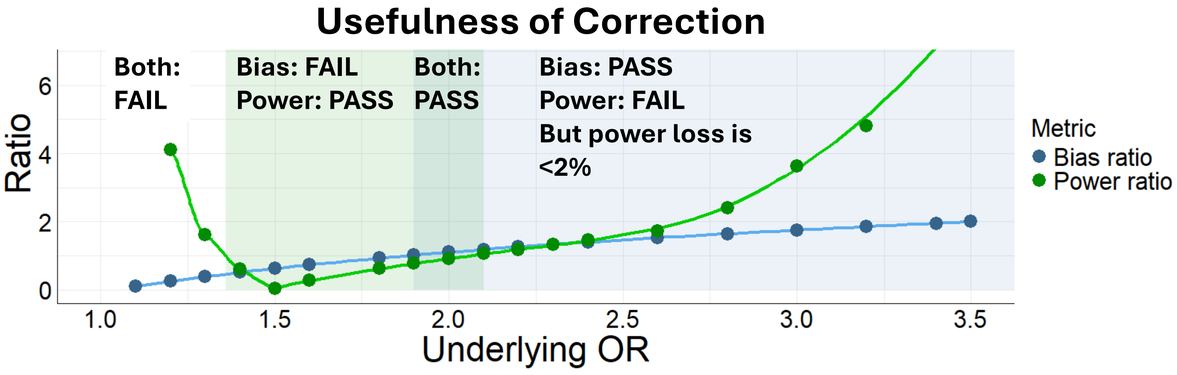
- Tradeoff between bias reduction and increased standard deviation (SD) impacts two measures, where β_1=ln(OR):

$$Bias\ Ratio = \frac{abs(Ave\beta_{1,pop} - Ave\beta_{1,mis}) - abs(Ave\beta_{1,pop} - Ave\beta_{1,corr})}{abs(SD_{corr} - SD_{mis})}$$

$$Power\ Ratio = \frac{abs(Power_{corr} - Power_{pop})}{abs(Power_{mis} - Power_{pop})}$$

PASS when: **Bias ratio > 1**, **Power ratio < 1**

Usefulness of Correction



In these simulations, correction reduced bias but increased variability, and its effect on power depended on the underlying OR. Bias reduction, changes in power, and false-positive rates should be considered together when evaluating the trade-off.