**SUPPLEMENTAL MATERIAL**

**Contents**

1. Worked Example
2. Algebraic Demonstration
3. Calculation of Exposure Weights Using Sensitivity and Specificity

**A. Worked Example**

In this Supplement, we demonstrate bias adjustment using a worked example. First, we artificially add biases to a hypothetical case-control study, and then show how to adjust for these biases. Numeric values were chosen to represent realistic scenarios but were not based on real data. Readers seeking algebraic proofs can find these in Section B.

We created a hypothetical population of 100,000 individuals with a dichotomous exposure (E), disease (D), and confounder (C), in which E causes D, and C causes both E and D (Table A1). We assume that variables are measured without error and have no missing data. This population is the “source population”. The confounder-adjusted Mantel-Haenszel odds ratio (OR) for the exposure-disease relationship is 2.79. This is the true OR in the population, which we take as causal absent other sources of bias.

We then designed a hypothetical case-control study in the source population, intending to select 100% of diseased individuals as cases and 10% of unaffected individuals as controls. However, we introduced selection bias, making selection (S; S=1 if participant is selected, S=0 if not) dependent on E and D: we selected 90% of exposed cases, 95% of non-exposed cases, 8.0% of exposed controls, and 9.9% of non-exposed controls. This is the “selected population” (Table A1). The confounder-adjusted OR is 3.34, representing effects of selection bias.

Next, we misclassified exposure differentially by D by applying sensitivities (Se) and specificities (Sp) of exposure classification to the selected population: Se(cases)=0.95, Se(controls)=0.90, Sp(cases)=0.98, and Sp(non-cases)=0.95. EM denotes misclassified exposure (Table A1). These are the “observed data”. The confounder-adjusted OR is 1.58, reflecting the combined effects of biases.

The observed data represent what investigators typically have to analyze, and are the data we will use to demonstrate the bias analysis step-by-step.

**Step 1: Estimate exposure misclassification weights**

We use the method of Lyles and Lin to estimate exposure misclassification weights using predictive values.

Because we have hypothetical data, both true (E) and misclassified (EM) exposure are known and we can calculate predictive values, p(E|EM), directly from the data (Table A2). (For real data, predictive values might be obtained from validation studies or educated guesses.) We estimated predictive values conditional on C so they could differ across strata of C, on D to allow for differential misclassification, and on S because the values are estimated among individuals in the study. For example, p(E=1| EM=1,C=1,D=1, S=1) is 254/272 and p(E=0| EM=1,C=1,D=1, S=1) is 18/272 (Table A2). These predictive values are the exposure weights for the analysis. In Section B, we show how to create exposure weights from Se and Sp instead.

**Step 2: Estimate selection weights**

We use the method of Hernán et al. to estimate inverse probability of selection weights (IPSW). First, we calculate selection probabilities conditional on E, D, and C (these might be conditional on EM instead if estimated from a population with exposure misclassification). Because we have hypothetical data, we calculate selection probabilities directly from Table A1. (In practice, selection probabilities might be estimated from participation rates or educated guesses.) For example, the selection probability p(S=1|E=1,D=1,C=1) is 265/299 and the IPSW is 299/265.

**Step 3: Multiple bias analysis**

We create a dataset in which each participant appears twice (Table A3). One participant copy is assigned as exposed (EA=1) and the other as unexposed (EA=0), because we do not know which is true. For example, the first 2 rows of Table A3 are the 272 individuals with C=1, D=1, and EM=1. One copy (row 1) is assigned EA=1 and the other (row 2), EA=0. From now on, we treat EA as the “true” exposure.

Next, we adjust for exposure misclassification. Each participant copy is assigned its weight (predictive value) which represents the probability that the value of EA is correct. Earlier, we calculated p(E=1|EM=1C=1,D=1,S=1)=254/272 and p(E=0|EM=1,C=1,D=1,S=1)=18/272, as weights for the participants with EA=1 and EA=0. The observed N in each row is multiplied by its weight to produce the misclassification-adjusted N. The weights sum to 1 within strata of C, D, and EM so that after applying the weights, the sample size is unchanged (e.g., in the first 2 rows of Table A3, the misclassification-adjusted N is 254+18=272 despite starting with 2 groups of 272 individuals).

Then we adjust for selection bias. EM is not needed anymore; it is deleted and the misclassification-adjusted N with identical values of C, D, and EA from Table A3 are collapsed (Table A4). The misclassification-adjusted N are multiplied by IPSWs. For example, we calculated the IPSW for p(S=1|E=1,D=1,C=1) as 299/265, which we apply to row 1.

The final N are the exposure misclassification- and selection bias-adjusted contingency table cell counts, which are used to calculate the OR. Standard errors should not be calculated from these data because they do not take into account variability added by estimating bias parameters. The bias analysis has recreated the original contingency table counts exactly (Table A1), successfully adjusting for multiple biases. The counts are identical because the exposure misclassification weights and IPSWs were calculated from hypothetical data. If weights are inaccurately estimated, results could differ from the truth.

We applied the exposure weights and IPSWs sequentially, but this can be done simultaneously by multiplying the exposure misclassification weight and IPSW to create a combined weight for each row in Table A3.

To perform this analysis in statistical software, one dataset copy is created with a new variable, EA=1. A second copy is created with EA=0. These are combined, creating a dataset with two participant copies. Each observation is assigned exposure and selection weights corresponding to its values of C, D, EM, and EA. The two weights are multiplied to create a combined weight. In the logistic regression model, EA is the exposure variable (instead of EM) and the combined weight is the analytic weight. Confounding is adjusted for by multivariable modeling, and the OR is the multiple bias-adjusted estimate of the association. The non-probabilistic multiple bias analysis can be implemented in any standard statistical software that allows specification of a weight variable in a logistic regression model. For example, in SAS (SAS Institute, Cary, NC), the code would follow this format in PROC LOGISTIC to adjust for exposure misclassification, selection bias, and confounding:

proc logistic data = dataset\_with\_participant\_copies;

model outcome (event = “1”) = assigned\_exposure confounders;

weight final\_weight;

run;

Note that the exposure variable in this code is the assigned exposure, not the measured (observed) exposure. The standard errors produced from the software do not take into account uncertainty added by estimation of the bias parameters and therefore should not be used.

Table A1. Contingency Tables for the Hypothetical Source Population, Selected Population, and Observed Data, Stratified by a Confounder.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | C = 1 | | C = 0 | | Adjusted ORa |
| Source population | E = 1 | E = 0 | E = 1 | E = 0 | 2.79 |
| D = 1 | 299 | 1,105 | 535 | 4,080 |  |
| D = 0 | 730 | 7,779 | 3,903 | 81,569 |  |
| Selected populationb | E = 1 | E = 0 | E = 1 | E = 0 | 3.34 |
| D = 1 | 265 | 1,042 | 476 | 3,902 |  |
| D = 0 | 58 | 774 | 293 | 7,987 |  |
| Observed datac | EM = 1 | EM = 0 | EM = 1 | EM = 0 | 1.58 |
| D = 1 | 272 | 1,035 | 523 | 3,855 |  |
| D = 0 | 99 | 733 | 688 | 7,592 |  |

Abbreviations: C, confounder; D, disease; E, true exposure; EM, misclassified exposure; OR, odds ratio.

a Mantel-Haenszel odds ratio adjusted for C.

b Addition of selection bias.

c Addition of exposure misclassification.

Table A2. Exposure Classification Table From the Hypothetical Selected Population, Stratified by Disease and Confounder.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | C = 1 | | | C = 0 | | |
|  | E = 1 | E = 0 | Total | E = 1 | E = 0 | Total |
| D = 1 |  |  |  |  |  |  |
| EM = 1 | 254 | 18 | 272 | 451 | 72 | 523 |
| EM = 0 | 11 | 1,024 | 1,035 | 25 | 3,830 | 3,855 |
| Total | 265 | 1,042 | 1,307 | 476 | 3,902 | 4,378 |
| D = 0 |  |  |  |  |  |  |
| EM = 1 | 52 | 47 | 99 | 262 | 426 | 688 |
| EM = 0 | 6 | 727 | 733 | 31 | 7,561 | 7,592 |
| Total | 58 | 774 | 832 | 293 | 7,987 | 8,280 |

Abbreviations: C, confounder; D, disease; E, true exposure; EM, misclassified exposure.

Table A3. Procedure to Adjust for Exposure Misclassification in the Hypothetical Study.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| C | D | EM | EA | Observed Na | p(E|C,D,EM,S)b | Misclassification-Adjusted Nc |
| 1 | 1 | 1 | 1 | 272 | 254/272 | 254 |
| 1 | 1 | 1 | 0 | 272 | 18/272 | 18 |
| 1 | 1 | 0 | 1 | 1,035 | 11/1,035 | 11 |
| 1 | 1 | 0 | 0 | 1,035 | 1,024/1,035 | 1,024 |
| 1 | 0 | 1 | 1 | 99 | 52/99 | 52 |
| 1 | 0 | 1 | 0 | 99 | 47/99 | 47 |
| 1 | 0 | 0 | 1 | 733 | 6/733 | 6 |
| 1 | 0 | 0 | 0 | 733 | 727/733 | 727 |
| 0 | 1 | 1 | 1 | 523 | 451/523 | 451 |
| 0 | 1 | 1 | 0 | 523 | 72/523 | 72 |
| 0 | 1 | 0 | 1 | 3,855 | 25/3,855 | 25 |
| 0 | 1 | 0 | 0 | 3,855 | 3,830/3,855 | 3,830 |
| 0 | 0 | 1 | 1 | 688 | 262/688 | 262 |
| 0 | 0 | 1 | 0 | 688 | 426/688 | 426 |
| 0 | 0 | 0 | 1 | 7,592 | 31/7,592 | 31 |
| 0 | 0 | 0 | 0 | 7,592 | 7,561/7,592 | 7,561 |

Abbreviations: C, confounder; D, disease; E, true exposure; EA, assigned exposure; EM, misclassified exposure; N, number of observations; p(E|C,D,EM,S), probability of E given C, D, and EM among those selected (S) into the study.

a Number of participants in the observed data given values of C, D, and EM (from Table A1). Each participant is entered into the analysis twice: once with EA = 1 and once with EA = 0.

b Predictive values calculated from the selected population (Table A2).

c Product of observed N and p(E|C,D,EM,S).

Table A4. Adjusting for Selection Bias in the Hypothetical Study.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| C | D | EA | Misclassification-Adjusted Na | IPSWb | Final Nc |
| 1 | 1 | 1 | 254 + 11 = 265 | 299/265 | 299 |
| 1 | 1 | 0 | 18 + 1,024 = 1,042 | 1,105/1,042 | 1,105 |
| 1 | 0 | 1 | 52 + 6 = 58 | 730/58 | 730 |
| 1 | 0 | 0 | 47 + 727 = 774 | 7,779/774 | 7,779 |
| 0 | 1 | 1 | 451 + 25 = 476 | 535/476 | 535 |
| 0 | 1 | 0 | 72 + 3,830 = 3,902 | 4,080/3,902 | 4,080 |
| 0 | 0 | 1 | 262 + 31 = 293 | 3,903/293 | 3,903 |
| 0 | 0 | 0 | 426 + 7,561 = 7,987 | 81,569/7,987 | 81,569 |

Abbreviations: C, confounder; D, disease; EA, assigned exposure; IPSW, inverse probability of selection weight; N, number of observations.

a Number of observations in strata of C, D, and EA (from Table A3, summed over misclassified exposure, EM).

b Inverse of selection probabilities (from Table A1).

c Product of misclassification-adjusted N and IPSW. When predictive value and selection proportions are known with certainty, the final N is equal to the number of individuals in the source population (Table A1). Outside of simulation, obtaining the same counts as the source population would be unlikely.**B. Algebraic Demonstration**

Here, we show algebraically that multiple bias analysis using a weighted analysis, when applied to observed data with exposure misclassification and selection bias, leads, on average, to the same joint distribution of exposure and disease as the source population when predictive values for exposure and selection probabilities are known. For demonstration purposes only, misclassification and selection bias are adjusted for sequentially; the same result would be obtained if the weights were applied simultaneously. Adjustment for confounding can also be incorporated, as shown in the main text.

Table B1 shows the joint distribution of true exposure (E) and disease (D) in the source and selected populations. We assume that the selected population was obtained by sampling from the source population with selection probabilities p(S)=πij (i=disease, j=exposure). Table B1 also shows the expected number of participants in the selected population, based on this selection.

We assume that exposure misclassification (EM) was then introduced by measuring exposure in the selected population with sensitivities and specificities Sei and Spi (Table B2). This process leads to the expected number of participants in the observed data, M1 (number with D=1) and M0 (number with D=0).

We now demonstrate that the expected cell counts after adjustment for exposure misclassification and selection bias using a weighted analysis in the observed data leads, on average, back to the cell counts of the source population contingency table when the predictive values for exposure classification and the selection probabilities are known.

The exposure classification table is shown in Table B3. The right marginal total is fixed at the observed number of participants in each cell of the contingency table and is equal to the expected number of participants in the selected population stratified by D and summed over E. The middle columns are the expected number of participants in the selected population with E=1 and E=0 for each category of D and EM, from which the predictive values (exposure misclassification weights) can be calculated as the cell-specific expected values divided by the marginal row totals.

Table B4 demonstrates the first step for multiple bias analysis: adjustment for exposure misclassification. Each participant is entered into the analysis twice: one copy of the participant is assigned to be exposed (EA=1) and the other to be unexposed (EA=0). Each observation is then multiplied (weighted) by its corresponding predictive value, p(E|D,EM,S), calculated from the exposure classification table (Table B3), to produce the expected values for the misclassification-adjusted counts (labeled “Misclassification Adjusted N” in Table B4).

Table B5 shows the misclassification-adjusted counts from Table B4 summed over EM; these expected counts are equal to the expected counts of the contingency table for the selected population. To adjust for selection bias, each observation is multiplied (weighted) by the inverse of its selection probability (inverse probability of selection weight).

After adjustment for both exposure misclassification and selection bias using the weighted analysis, the final expected counts are equal to the contingency table counts from the source population.

A nearly identical argument shows that the source population counts can be replicated exactly if the exact predictive proportions and exact selection proportions for these populations are known (we define “predictive proportion positive” as the proportion of those measured as exposed who are truly exposed, with a similar definition for the unexposed).

Table B1. The Source and Selected Populations.

|  |  |  |  |
| --- | --- | --- | --- |
|  | E = 1 | E = 0 | |
| Source Population |  | |  |
| D = 1 | a | | b |
| D = 0 | c | | d |
| Selected Population |  | |  |
| D = 1 | π11a | | π10b |
| D = 0 | π01c | | π00d |

Abbreviations: D, disease; E, true exposure.

Table B2. Observed Data With Exposure Misclassification and Selection Bias, Representing Data Usually Available in Epidemiologic Studies.

|  |  |  |  |
| --- | --- | --- | --- |
| Observed Data | EM = 1 | EM = 0 | Total |
| D = 1 | aʹ = (Se1) π11a + (1-Sp1) π10b | bʹ = (1-Se1) π11a + (Sp1) π10b | M1 |
| D = 0 | cʹ = (Se0) π01c + (1-Sp0) π00d | dʹ = (1-Se0) π01c + (Sp0) π00d | M0 |

Abbreviations: D, disease; E, true exposure; EM, misclassified exposure; i = disease status (1 = yes, 0 = no); Sei, sensitivity; Spi, specificity.

Table B3. Exposure Classification Table Stratified by Disease.

|  |  |  |  |
| --- | --- | --- | --- |
|  | E = 1 | E = 0 | Total |
| D = 1 |  |  |  |
| EM = 1 | (Se1) π11a | (1-Sp1) π10b | aʹ = (Se1) π11a + (1-Sp1) π10b |
| EM = 0 | (1-Se1) π11a | (Sp1) π10b | bʹ = (1-Se1) π11a + (Sp1) π10b |
| Total | π11a | π10b | M1 |
| D = 0 |  |  |  |
| EM = 1 | (Se0) π01c | (1-Sp0) π00d | cʹ = (Se0) π01c + (1-Sp0) π00d |
| EM = 0 | (1-Se0) π01c | (Sp0) π00d | dʹ = (1-Se0) π01c + (Sp0) π00d |
| Total | π01c | π00d | M0 |

Abbreviations: D, disease; E, true exposure; EM, misclassified exposure; i, disease status (1 = yes, 0 – no); Sei, sensitivity; Spi, specificity.

Table B4. Adjustment for Exposure Misclassification

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| D | EM | EA | Observed N | p(E|D,EM,S) | Misclassification Adjusted Na |
| 1 | 1 | 1 | aʹ | (Se1) π11a/aʹ | (Se1) π11a |
| 1 | 1 | 0 | aʹ | (1-Sp1) π10b/aʹ | (1-Sp1) π10b |
| 1 | 0 | 1 | bʹ | (1-Se1) π11a/bʹ | (1-Se1) π11a |
| 1 | 0 | 0 | bʹ | (Sp1) π10b/bʹ | (Sp1) π10b |
| 0 | 1 | 1 | cʹ | (Se0) π01c/cʹ | (Se0) π01c |
| 0 | 1 | 0 | cʹ | (1-Sp0) π00d/cʹ | (1-Sp0) π00d |
| 0 | 0 | 1 | dʹ | (1-Se0) π01c/dʹ | (1-Se0) π01c |
| 0 | 0 | 0 | dʹ | (Sp0) π00d/dʹ | (Sp0) π00d |

Abbreviations: D, disease; E, exposure; EM, misclassified exposure; N, number of observations; p(E|D,EM,S), probability of true exposure, given disease, misclassified exposure, and selection into the study; Sei, sensitivity; Spi, specificity.

a Product of observed N and p(E|D,EM,S).

Table B5. Adjustment for Selection Bias in Addition to Exposure Misclassification

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| D | EA | Misclassification Adjusted Na | IPSW | Final Nb |
| 1 | 1 | (Se1) π11a + (1-Se1) π11a = π11a | 1/ π11 | π11a/π11 = a |
| 1 | 0 | (1-Sp1) π10b + (Sp1) π10b = π10b | 1/ π10 | π10b/π10 = b |
| 0 | 1 | (Se0) π01c + (1-Se0) π01c = π01c | 1/ π01 | π01c/π01 = c |
| 0 | 0 | (1-Sp0) π00d + (Sp0) π00d = π00d | 1/ π00 | π00d/π00 = d |

Abbreviations: D, disease; EA, assigned exposure; IPSW, inverse probability of selection weight; N, number of observations; Se, sensitivity; Sp, specificity.

a From Table B4, summed over EM.

b Product of misclassification adjusted N and IPSW.

**C. Calculation of Exposure Weights Using Sensitivity and Specificity**

If predictive values for exposure are unavailable but values of Se and Sp are available for cases and controls, these can be used to calculate the weights. Here, we derive these weights and show that they are equivalent to using predictive values in the weighted analysis. We use the “a” cell (D=1, E=1) as demonstration, but similar arguments can be used to derive weights for the other cells.

Positive predictive value (PPV), the exposure weight for the aʹ cell, can be expressed in terms of Se and Sp:

PPV =

Because p(E), prevalence of E, is unknown in the observed data, we must express it in terms of known quantities. To do this, we first express π11a in terms of known quantities. Rearranging from Table B3:

aʹ = Se(π11a) + (1-Sp) π10b

a’ = Se(π11a) + (1-Sp)(M1- π11a)

a’ = Se(π11a) + M1 – M1Sp + Sp(π11a) - π11a

a’ = (π11a)(Se + Sp – 1) + M1(1 – Sp)

π11a =

Prevalence of E becomes:

p(E) =

=

=

We now substitute this quantity into the equation for PPV and rearrange:

PPV =

=

=

=

=

This quantity can be used as the exposure weight. From Table B4, the misclassification adjusted N = Se1(π11a). This is equivalent to aʹ PPV (i.e., multiplying the misclassified cell count by the weight):

aʹ PPV =

=

= Se(π11a)

Similar derivations show that Se and Sp can be substituted for other predictive values for the weights (Table C1).

Table C1. Exposure Weights Using Sensitivity and Specificity.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| D | EM | EA | Observed N | Exposure Weight | Misclassification-Adjusted N |
| 1 | 1 | 1 | aʹ |  | (Se1)π11a |
| 1 | 1 | 0 | aʹ |  | (1-Sp1) π10b |
| 1 | 0 | 1 | bʹ |  | (1-Se1) π11a |
| 1 | 0 | 0 | bʹ |  | (Sp1) π10b |
| 0 | 1 | 1 | cʹ |  | (Se0) π01c |
| 0 | 1 | 0 | cʹ |  | (1-Sp0) π00d |
| 0 | 0 | 1 | dʹ |  | (1-Se0) π01c |
| 0 | 0 | 0 | dʹ |  | (Sp0) π00d |

Abbreviations: D, disease; EA, assigned exposure; EM, misclassified exposure; N, number of observations; Sei, sensitivity; Spi, specificity.