BIOS 736, Fall 2023: Notes for Class # 2

WHAT IS REQUIRED FOR AN ACTUAL MEASUREMENT ERROR OR MISCLASSIFICATION ADJUSTMENT?

First, let's return to exposure misclassification

Recall the scenario for a case-control study (independent samples): true exposure ("gold standard")

True Cell Probabilities

	X	
D	1	0
1	π_1	$1-\pi_1$
0	π_0	$1-\pi_0$

$\Rightarrow$

"True" Cell Counts

	X	
D	1	0
1	a	b
0	c	d

"Observed" Cell Probabilities

	W	
D	1	0
1	π_1^*	$1-\pi_1^*$
0	π_0^*	$1-\pi_0^*$

$\Rightarrow$

Observed Cell Counts (error-prone version of X)

	W	
D	1	0
1	A	B
0	C	D

Again, $\pi_1 = \Pr(X=1 \mid D=1)$ prob of exposure among cases

$\pi_0 = \Pr(X=1 \mid D=0)$ prob of exposure among controls

Parameter of interest: $\boxed{OR = \pi_1(1-\pi_0)/[\,\pi_0(1-\pi_1)]}$

Note that it is $\pi_1^* = \Pr(W=1 \mid D=1)$ and $\pi_0^* = \Pr(W=1 \mid D=0)$ that are directly estimable from the observed (D, W) table, and these estimates will generally be biased for π_1 and π_0 due to misclassification

Allowing for possible differential misclassification, recall these definitions:

Sensitivities: $SE_d = \Pr(W=1 \mid X=1, D=d)$

Specificities: $SP_d = \Pr(W=0 \mid X=0, D=d)$ $\quad (d=0,1)$

(4) parameters.

If we knew these values, we could directly implement a correction of the "naïve" π_1^* and π_0^* estimates:

$$\pi_d^* = \Pr(W=1 \mid D=d) = \Pr(W=1, X=1 \mid D=d) + \Pr(W=1, X=0 \mid D=d)$$

$$= \Pr(W=1 \mid X=1, D=d) \times \Pr(X=1 \mid D=d)$$

$$+ \Pr(W=1 \mid X=0, D=d) \times \Pr(X=0 \mid D=d)$$

$$= SE_d\pi_d + (1-SP_d)(1-\pi_d) \qquad (d=0,1)$$

By solving for π_d, an equivalent identity becomes: (from W vs. D table)

$$\pi_d = \frac{\hat{\pi}_d^* + SP_d - 1}{SE_d + SP_d - 1} \qquad (d=0, 1)$$

- NOTE: These two identities are the basis of the so-called "matrix method", which has roots in classic papers on misclassification (e.g., Bross 1954 *Biometrics*; Barron 1977 *Biometrics*)
- An important point (e.g., Bross 1954) is that in non-differential case, H_0: OR $= 1 \Rightarrow$ OR* $= 1$. Thus, standard inference remains valid based only on the (D,W) table!

Test of OR=1 based on W vs. D table is valid

$SE_1 = SE_0 = \Pr(W=1 \mid X=1)$

$SP_1 = SP_0 = \Pr(W=0 \mid X=0)$

ASIDE: Why do they call it the "matrix method" (e.g., Barron, 1977)?

In this case-control setting, we can write:

$$\begin{pmatrix} \pi_1^* \\ 1-\pi_1^* \\ \pi_0^* \\ 1-\pi_0^* \end{pmatrix} = \begin{pmatrix} SE_1 & 1-SP_1 & 0 & 0 \\ 1-SE_1 & SP_1 & 0 & 0 \\ 0 & 0 & SE_0 & 1-SP_0 \\ 0 & 0 & 1-SE_0 & SP_0 \end{pmatrix} \begin{pmatrix} \pi_1 \\ 1-\pi_1 \\ \pi_0 \\ 1-\pi_0 \end{pmatrix}$$

$$\underset{\sim}{\pi}^* \qquad \underset{\sim}{A} \qquad \underset{\sim}{\pi}$$

$$\Rightarrow \quad \underset{\sim}{\pi}^* = \underset{\sim}{A}\,\underset{\sim}{\pi}$$

$$\Rightarrow \quad \boxed{\underset{\sim}{\pi} = \underset{\sim}{A}^{-1}\,\underset{\sim}{\pi}^*}$$

Note: If we had no data on which to base estimates of SE_d and SP_d, a reasonable possibility might be to conduct a "sensitivity" analysis:

⇒ Choose plausible combinations of (SE_d, SP_d), and compute the corresponding values of π_d after substituting the observed-data estimates of π_d^* (from D, W table) into the previous identity on pg. 2. Note: $\hat{\pi}_1^* = \frac{A}{A+B}$, $\hat{\pi}_0^* = \frac{C}{C+D}$

⇒ Provides an idea of the range of possible OR values (but, should consider variability due to uncertainty in the $\hat{\pi}_d^*$'s)

See Lyles and Lin (2010) for more examples of sensitivity/uncertainty analysis (in logistic regression)

But let's assume we have external validation data, of the following typical form:

error-prone

"gold standard"

External validation data cell counts

	W	
X	1	0
1	E	F
0	G	H

$\hat{SE} = \hat{P}_r(W=1 \mid X=1)$

$\hat{SP} = \hat{P}_r(W=0 \mid X=0)$

⇒ $\hat{SE} = \frac{E}{E+F}$, $\hat{SP} = \frac{H}{G+H}$

$\hat{Var}(\hat{SE}) = \frac{\hat{SE}(1-\hat{SE})}{E+F}$, $\hat{Var}(\hat{SP}) = \frac{\hat{SP}(1-\hat{SP})}{G+H}$, $\hat{Cov}(\hat{SE}, \hat{SP}) \cong 0$

Note: structure of external validation data forces us to assume non-differentiality!

Now we can get an actual "corrected" OR estimate:

$$\hat{\pi}_d = \frac{\hat{\pi}_d^* + \hat{SP}_d - 1}{\hat{SE}_d + \hat{SP}_d - 1} \quad (d=0, 1) \quad \Rightarrow \quad \hat{OR} = \frac{\hat{\pi}_1(1-\hat{\pi}_0)}{\hat{\pi}_0(1-\hat{\pi}_1)}$$

And, we could obtain an estimate of $\text{Var}[\ln(\hat{OR})]$ using (you guessed it) the delta method

Review origin and application of delta method here:

$$\ln\left(\hat{OR}\right) = \ln\left\{\frac{\hat{\pi}_1(1-\hat{\pi}_0)}{\hat{\pi}_0(1-\hat{\pi}_1)}\right\}$$

$$\Rightarrow \text{var}\left[\ln(\hat{OR})\right] = \text{var}\left[\ln\left(\frac{\hat{\pi}_1}{1-\hat{\pi}_1}\right) - \ln\left(\frac{\hat{\pi}_0}{1-\hat{\pi}_0}\right)\right]$$

$$= \sum_{d=0}^{1} \text{var}\left[\ln\left(\frac{\hat{\pi}_d}{1-\hat{\pi}_d}\right)\right] - 2\,\text{cov}\left[\downarrow, \downarrow\right]$$

Univariate Δ-method results:

If $Y = g(X)$

$$\Rightarrow \text{var}(Y) \doteq \left(\frac{dg}{dx}\Big|_{x=\mu_x}\right)^2 \text{var}(X)$$

Also, if $Y_1 = g_1(X_1)$ and $Y_2 = g_2(X_2)$

$$\Rightarrow \text{cov}(Y_1, Y_2) = \text{cov}\left[g_1(X_1), g_2(X_2)\right]$$

$$\doteq \left(\frac{dg_1}{dx_1}\Big|_{x_1=\mu_{x_1}}\right)\left(\frac{dg_2}{dx_2}\Big|_{x_2=\mu_{x_2}}\right)\text{cov}(X_1, X_2)$$

Let's use these results:

$$g(\hat{\pi}_d) = \ln\left[\hat{\pi}_d(1-\hat{\pi}_d)^{-1}\right], \quad d = 0, 1$$

$$\Rightarrow \frac{dg}{d\pi_d} = \frac{1}{\pi_d(1-\pi_d)}$$

$$\Rightarrow \text{var}\left[\ln(\widehat{OR})\right] \doteq \sum_{d=0}^{1} \frac{\text{var}(\hat{\pi}_d)}{\left[\hat{\pi}_d(1-\hat{\pi}_d)\right]^2} - \frac{2\,\text{cov}(\hat{\pi}_1, \hat{\pi}_0)}{\hat{\pi}_1(1-\hat{\pi}_1)\hat{\pi}_0(1-\hat{\pi}_0)}$$

* To get the final result, use multivariate Δ-method: to get $\hat{\text{var}}(\hat{\pi}_d)$ and $\text{cov}(\hat{\pi}_1, \hat{\pi}_0)$

$$\hat{\pi}_d = g\left(\hat{\pi}_d^*, \underbrace{SE, SP}_{\text{external validities}}\right)$$

($\hat{\pi}_d^*$: main study)

* $$\boxed{\hat{\text{var}}(\hat{\pi}_d) \doteq \underset{\sim}{\hat{D}}_d \underset{\sim}{\hat{\Sigma}} \underset{\sim}{\hat{D}}_d'}$$

where $\underset{\sim}{\hat{D}}_d$ is estimated vector of 1st derivatives of g w/respect π_d^*, SE, SP.

Also,

$$\boxed{\hat{\text{cov}}(\hat{\pi}_1, \hat{\pi}_0) = \underset{\sim}{\hat{D}}_1 \underset{\sim}{\hat{\Sigma}} \underset{\sim}{\hat{D}}_0'}$$

where $\underset{\sim}{\hat{\Sigma}} = \hat{\text{var}}\begin{pmatrix} \hat{\pi}_d^* \\ \hat{SE} \\ \hat{SP} \end{pmatrix}$ (3×3)

$\Rightarrow$ Now can get $\hat{SE}\left[\ln(\widehat{OR})\right]$, ~ 95% CI

Now, what if the validation data are internal:

⇒ X measured (in addition to D and W) on a random subsample of your subjects

Let's consider the pros/cons of external vs. internal validation designs:

Validation Study Characteristics

	Advantages	Disadvantages
External	CHEAP, EASY	MAY NOT BE AVAILABLE AND/OR "TRANSPORTABLE" ; NOT STATISTICALLY EFFICIENT ; LIKELY REQUIRES NON-DIFFERENTIAL ERROR ASSUMPTION
Internal	CAN BE COSTLY, TIME CONSUMING	SOLVES "TRANSPORTABILITY" PROBLEM ; STATISTICALLY EFFICIENT ; NON-DIFFERENTIALITY NOT REQUIRED

⇒ Different data layout than for the external case:

				Internal Validation subsample				
					D=1		D=0	
Main study sample				W	X=1	X=0	X=1	X=0
D	W=1	W=0						
1	n_{11}	n_{12}		1	n_{13}	n_{14}	n_{03}	n_{04}
0	n_{01}	n_{02}		0	n_{15}	n_{16}	n_{05}	n_{06}

n_{dj} $(d=0,1)$, $j=1,\ldots 6$

* For this study design, we could specify and maximize the likelihood function accounting for all 12 types of observations (12 different cell counts)

ex) What is the likelihood contribution for each of the n_{11} subjects in the main study with (D=1, W=1)?

$\pi_1^* = \Pr(W=1 \mid D=1) = SE_1\pi_1 + (1-SP_1)(1-\pi_1)$ (see pg. 2)

↓ could do for all 12 cell counts

- Note: formulating the likelihood based on such "matrix-method" identities [i.e., parameterizing in terms of (π_d, SE_d, SP_d)] leads to a function that can be maximized numerically, but does not appear to yield closed-form MLEs.

In this case, an alternative parameterization [in terms of (π^*_d, PPV_d, NPV_d)] turns out to be useful

- Can show (Lyles 2002 *Biometrics*) that the ML estimates for this setting are identical to the so-called "inverse matrix" estimators (Marshall 1990 *J Clinical Epidemiology*)

The "inverse matrix" method is based on a similar but different identity to that underlying the "matrix" method:

$PPV_1 = \Pr(X=1 \mid W=1, D=1)$
$PPV_0 = \Pr(X=1 \mid W=1, D=0)$

$$\pi_d = \Pr(X=1 \mid D=d) = \Pr(X=1, W=1 \mid D=d) + \Pr(X=1, W=0 \mid D=d)$$

$$= \Pr(X=1 \mid W=1, D=d) \times \Pr(W=1 \mid D=d)$$

$$+ \Pr(X=1 \mid W=0, D=d) \times \Pr(W=0 \mid D=d)$$

$$= PPV_d\pi_d^* + (1-NPV_d)(1-\pi_d^*) \qquad (d=0,1)$$

where $PPV_d = \Pr(X=1 \mid W=1, D=d)$ and $NPV_d = \Pr(X=1 \mid W=0, D=d)$ are the "positive and negative predictive values"

Let's enumerate the likelihood contributions for each of the separate types of observations based on this second parameterization: $(D=1)$

# obs	obs. type	Probability (likelihood contributes)
n_{11}	$W=1, D=1$	$Pr(W=1 \mid D=1) = \pi_1^*$
n_{12}	$W=0, D=1$	$(1-\pi_1^*)$
n_{13}	$W=1, X=1, D=1$	$PPV_1 \pi_1^*$
n_{14}	$W=1, X=0, D=1$	$(1-PPV_1)\pi_1^*$
n_{15}	$W=0, X=1, D=1$	$(1-NPV_1)(1-\pi_1^*)$
n_{16}	$W=0, X=0, D=1$	$NPV_1 (1-\pi_1^*)$

question (sum to 1)?

$\downarrow$

$NPV_1 = Pr(X=0 \mid W=0, D=1)$

$\Rightarrow$ (up to a constant,

$\log \mathcal{L}$ is

$$\ell = \sum_{d=0}^{1} \{ n_{d1} \ln(\pi_d^*) + n_{d2} \ln(1-\pi_d^*) + \dots (4 \text{ terms}) \}$$

$\Rightarrow$ take $\begin{cases} \frac{\partial \ell}{\partial PPV_d} = \dots = 0 \\ \frac{\partial \ell}{\partial NPV_d} = \dots = 0 \\ \partial \ell / \partial \pi_d^* = \end{cases}$

Can take those 3 derivatives easily if you want (results on next pg.)

Based on the study design in the previous table and assuming differential misclassification, the MLE's for the parameters involved in the "inverse matrix method" identity are:

$$\hat{\pi}_d{}^* = \frac{n_{d1}+n_{d3}+n_{d4}}{n_d}\ ,\ \hat{PPV}_d = \frac{n_{d3}}{n_{d3}+n_{d4}}\ ,\ \text{and}\ \hat{NPV}_d = \frac{n_{d6}}{n_{d5}+n_{d6}}\quad (d=0,1),$$

where $n_d = n_{d1}+n_{d2}+n_{d3}+n_{d4}+n_{d5}+n_{d6}$

* We can then use the preceding inverse matrix identity to get the MLEs for π_d and for the OR, and again can use the delta method (or the observed information matrix) to estimate $\text{Var}[\ln(\hat{OR})]$:

$$\hat{\pi}_d = \hat{PPV}_d\hat{\pi}_d{}^* + (1-\hat{NPV}_d)(1-\hat{\pi}_d{}^*)$$

* The MLE for π_d is same as Marshall's "inverse matrix" estimator.

* In differential case, $\hat{\pi}_1 \perp \hat{\pi}_0$ → because of case-control sampling design

$$\Rightarrow \hat{\text{var}}\left[\ln(\hat{OR}_{ML})\right] \doteq \sum_{d=0}^{1} \frac{\hat{\text{var}}(\hat{\pi}_d)}{[\hat{\pi}_d(1-\hat{\pi}_d)]^2}$$

↓ Marshall (1990), Morrissey & Spiegelman (1999), Lyles (2002) → (HW 1)

↓ no $\text{cov}(\hat{\pi}_1, \hat{\pi}_0)$ here.

Interesting note: For this main study / internal validation design, there is no known closed-form MLE for the OR when the misclassification is assumed to be *non-differential*

⇒ Obtain MLE by numerical maximization of the likelihood function

OR

⇒ Inverse variance-weighted ln(OR) estimators – weighting the estimators for ln(OR) based on i) the internal validation (D,X) data and ii) the main study + validation study data where the latter are treated "as external"

(e.g., Greenland 1987 *Stats in Medicine*; Thurston et al. 2005 *J Stat Planning and Inference;* Lyles et al. 2007 *Epidemiology*)

Basic idea of Greenland's (1987) closed-form "weighted" estimator:

$$\hat{\theta} = \hat{w}\,\hat{\theta}_I + (1-\hat{w})\,\hat{\theta}_E$$

almost as efficient as MLE!

$\hat{w}$ is an inverse-variance weight

$\hat{\theta}$ ↓ $\log(\hat{OR})$

$\hat{\theta}_I = \ln(\hat{OR})$ based only on (D, X) pairs from internal validation sample

$\hat{\theta}_E = \ln(\hat{OR})$ based on (D, W) pairs in the main study, combined with the (X, W) pairs in the validation sample, treating it like an external validation sample.

(see Hw # 1)

Real Data Example

Consider the following cell counts from a case-control study of SIDS, where exposure (X) is maternal use of antibiotics during pregnancy:

Main study sample				Validation subsample				
					D=1		D=0	
D	W=1	W=0		W	X=1	X=0	X=1	X=0
1	122	442		1	29	22	21	12
0	101	479		0	17	143	16	168

*Refs: (Drews et al. 1990 *Int J Epidemiol*; Greenland 1988 *Stats in Med*; Marshall 1990 *J Clin Epidemiol*; Morrissey and Spiegelman 1999 *Biometrics*)

Note: "Naïve" OR estimate is $\frac{122 \times 479}{442 \times 101} = 1.31$

$\Rightarrow$ "Naïve" : $\hat{\ln(OR)} = 0.269$, $\hat{SE}[\ln(OR)] = 0.150$

usual Woolf's method

$$\sqrt{\frac{1}{A} + \frac{1}{B} + \frac{1}{C} + \frac{1}{D}}$$

In contrast, the MLE ("inverse matrix") estimator accouting for (assumed differential) misclassification is based on:

$$\hat{\pi}_1^* = \frac{122+29+22}{775} = 0.223 \ , \ \hat{PPV}_1 = \frac{29}{29+22} = 0.568 \ , \ \hat{NPV}_1 = \frac{143}{17+143} = 0.894 \ ,$$

$$\hat{\pi}_0^* = \frac{101+21+12}{797} = 0.168 \ , \ \hat{PPV}_0 = \frac{21}{21+12} = 0.636 \ , \ \hat{NPV}_0 = \frac{168}{16+168} = 0.913$$

$$\Rightarrow \quad \hat{\pi}_1 = 0.568 \times 0.223 \ + \ (1-0.894) \times (1-0.223) \ = \ 0.209$$

$$\hat{\pi}_0 = 0.636 \times 0.168 \ + \ (1-0.913) \times (1-0.168) \ = \ 0.179$$

$$\Rightarrow \quad \hat{OR} = 1.21 \ , \ \ln(\hat{OR}) = 0.192$$

vs. .150

adjusted (vs. 1.31 unadjusted)

Also get $SE[\ln(\hat{OR})] = 0.221$, via delta method or observed information matrix
Likelihood ratio test for nondifferentiality can also be conducted (p=0.14 here)

w/ Carey Drews-Botsch

(See, e.g., Lyles et al. 2007 *Epidemiology* for further analysis of this example)

used variations on Greenland's wtd avg. estimator

Allows using both internal and external validation data together, and testing for "transportability"

NOTE: How could one extend the "matrix method" idea to the cross-sectional study design case, with both X and Y misclassified?

As in Barron (1977), interested in OR based on 2x2 table, with cross-sectional sampling.

Y \ X	1	0
1	π_{11}	π_{01}
0	π_{10}	π_{00}

X = true exposure
Y = true outcome

$\pi_{xy} = Pr(X=x, Y=y)$
$(x, y) = (0, 1)$

Let X^* and Y^* be misclassified versions

Y^* \ X^*	1	0
1	π_{11}^*	π_{01}^*
0	π_{10}^*	π_{00}^*

$\pi_{xy}^* = Pr(X^*=x, Y^*=y)$
$(x, y) = (0, 1)$

Note:

$$\pi_{xy}^* = \sum_{i=0}^{1}\sum_{j=0}^{1} Pr(X^*=x, Y^*=y, X=i, Y=j)$$

Suppose make these assumptions:

a) $Pr(X^* | X, Y^*, Y) = Pr(X^* | X) = SE_X$

b) $Pr(Y^* | Y, X^*, X) = Pr(Y^* | Y) = SE_Y$

"independent and nondifferential misclassification".

Note, 1st term (of 4) for π_{11}^* is

$$Pr(x^*=1, Y^*=1, x=1, Y=1)$$

$$= Pr(x^*=1 \mid x=1, Y^*=1, Y=1)\, Pr(Y^*=1 \mid Y=1, x=1) \times \pi_{11}$$

$$= SE_x SE_y \pi_{11}$$

Similarly, other 3 terms: $SE_x(1-SP_y)\pi_{10}$

$(1-SP_x)SE_y\pi_{01}$

$(1-SP_x)(1-SP_y)\pi_{00}$

$\Rightarrow \pi_{11}^*$ is sum of these 4 terms.

Barron completed this exercise and wrote:

let $\underset{\sim}{\pi}^* = \begin{pmatrix} \pi_{11}^* & \pi_{10}^* \\ \pi_{01}^* & \pi_{00}^* \end{pmatrix}$, $\underset{\sim}{\pi} = \begin{pmatrix} \pi_{11} & \pi_{10} \\ \pi_{01} & \pi_{00} \end{pmatrix}$

$$\Rightarrow \underset{\sim}{\pi}^* = \underset{\sim}{A}' \underset{\sim}{\pi} \underset{\sim}{B}$$

$$= \begin{pmatrix} SE_x & 1-SP_x \\ 1-SE_x & SP_x \end{pmatrix} \begin{pmatrix} \pi_{11} & \pi_{10} \\ \pi_{01} & \pi_{00} \end{pmatrix} \begin{pmatrix} SE_y & 1-SE_y \\ 1-SP_y & SP_y \end{pmatrix}$$

$\underset{\sim}{A}'$ $\qquad$ $\underset{\sim}{\pi}$ $\qquad$ $\underset{\sim}{B}$

"matrix method"!

Tang et al. (2013) extended matrix to relax assumptions

$$\underset{\sim}{\pi} = (A')^{-1} \underset{\sim}{\pi}^* \underset{\sim}{B}^{-1}$$