Dealing with Misclassification in the Results, Not the Discussion Section: Introduction to Quantitative Bias Analysis

Matthew Fox
Departments of Epidemiology and Global Health
Boston University, USA
 mfox@bu.edu
 @ProfMattFox
 Free Associations Podcast:
www.pophealthex.org/FA

BOSTON UNIVERSITY

Essay

Why Most Published Research Findings Are False

John P. A. Ioannidis

Summary

There is increasing concern that most current published research findings are false. The probability that a research claim is true may depend on study power and bias, the number of other studies on the same question, and, importantly, the ratio

factors that influence this problem and some corollaries thereof.

Modeling the Framework for False Positive Findings

Several methodologists have pointed out [9–11] that the high rate of nonreplication (lack of

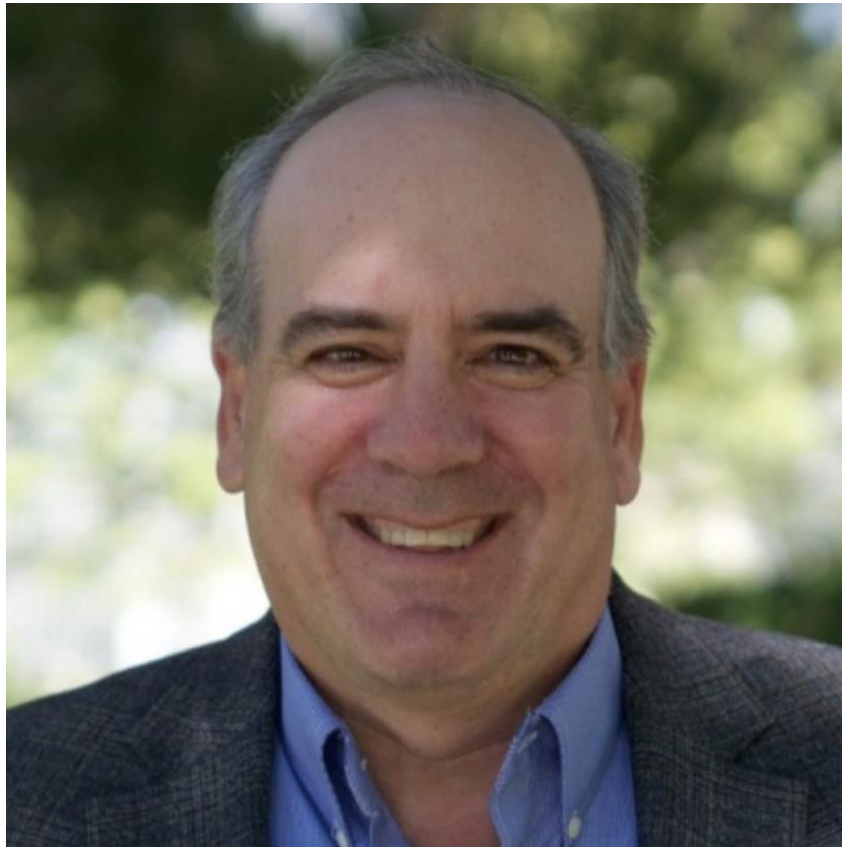
is characteristic of th
vary a lot depending
field targets highly li
or searches for only
true relationships an
and millions of hypo
be postulated. Let us
for computational si

How Do We Typically Work?

- **Conduct a research study using an appropriate study design**
  - Attempt to minimize bias in design, adjust for confounding
- **Calculate a measure of association**
  - Relative risk, risk difference
  - Calculate a summary of random error (e.g. pvalue or 95% CI)
- **Interpret results based on the random error alone**
- **Then think about bias**
  - Comment on it in our discussion sections

Feel Free to Ignore the Bias

- **Goodman S. Introduction to Bayesian Methods I: measuring the strength of evidence. Clinical Trials 2005; 2: 282-90**
  - "Let us remind ourselves what they have delivered into our laps. Here is a list of things that have been identified as cancer risks: electric razors; broken arms (but only in women); fluorescent lights; allergies; breeding reindeer; being a waiter; owning a pet bird; being short; being tall; and hot dogs. And, in case anyone is feeling safe - having a refrigerator [5]. We are apparently all at risk. These results were not produced by Bayesian methods."



BU

Premise

- **If:**
  - The objective of etiologic epidemiologic research is to obtain a **valid** and **precise** measure of the effect of an exposure on the occurrence of a disease
- **Then:**
  - Epidemiologists have an obligation to quantify how far from the objective their estimate might fall

Valid and precise

Not valid, precise

Valid, not precise

Not valid, not precise

What is the Role of QBA?

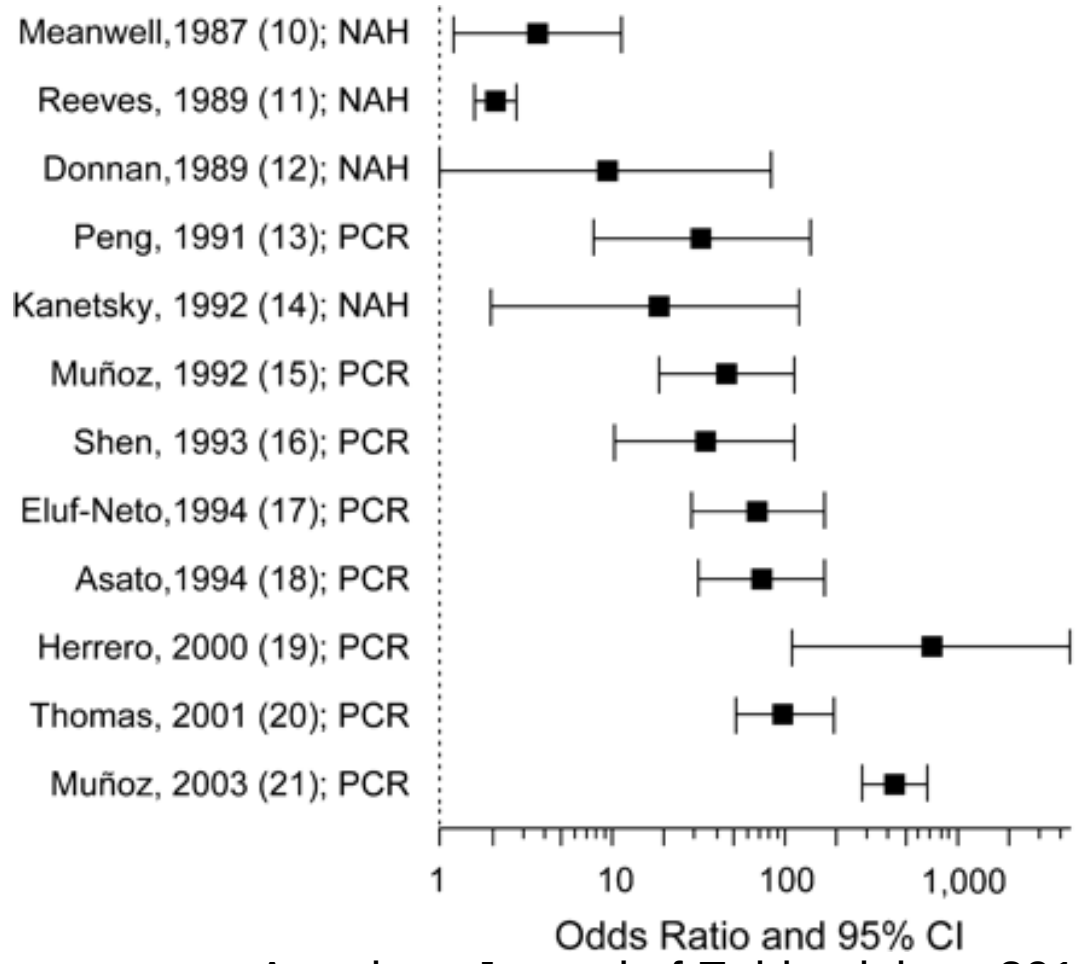
- **Way to relate estimates of bias and observed data to the true data**
- **Allows us to go beyond our assumptions about how bias behaves and quantify it**
- **Systematic approach to quantifying the impact of systematic error in terms of:**
  - Direction
  - Magnitude
  - Uncertainty

BU

QBA for misclassification

BOSTON UNIVERSITY

HPV and cervical cancer as HPV tests improved



Odds ratios for the association between human papillomavirus (HPV) infection and invasive cervical cancer risk in successive molecular epidemiologic studies

American Journal of Epidemiology 2010 171(2):164-168

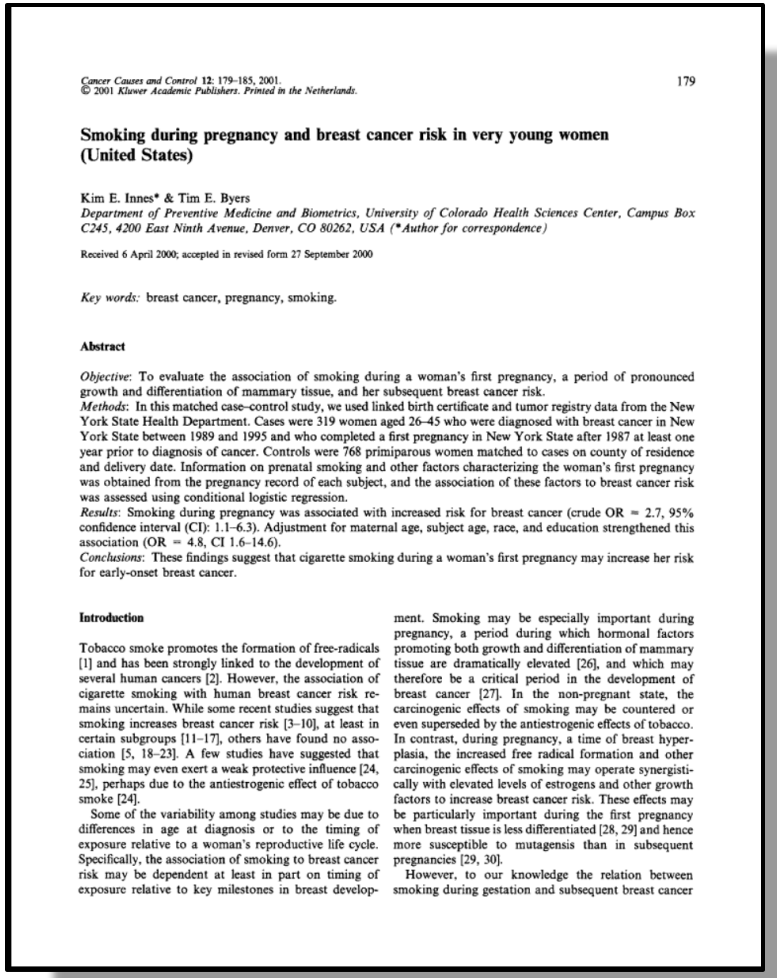
Exposure Misclassification

Motivating Example

BOSTON UNIVERSITY

Example Background

- **Research question:**
  - To investigate the association between smoking during pregnancy and breast cancer
- **Innes and Byers, 2001:**
  - Women who smoked during pregnancy had almost five times the risk of breast cancer as women who did not smoke during pregnancy (OR: 4.8; 95% CI: 1.6–14.6)
  - Wanted to replicate with similar design
- **Strengths of study**
  - Minimal selection bias, included all eligible cases
  - No recall bias because exposure and covariates assessed by review of birth certificates



Cancer Causes and Control 12: 179–185, 2001.

Smoking during pregnancy and breast cancer risk in very young women (United States)

Kim E. Innes* & Tim E. Byers
*Department of Preventive Medicine and Biometrics, University of Colorado Health Sciences Center, Campus Box C245, 4200 East Ninth Avenue, Denver, CO 80262, USA (*Author for correspondence)*

Received 6 April 2000; accepted in revised form 27 September 2000

Key words: breast cancer, pregnancy, smoking.

Abstract

Objective: To evaluate the association of smoking during a woman's first pregnancy, a period of pronounced growth and differentiation of mammary tissue, and her subsequent breast cancer risk.
Methods: In this matched case-control study, we used linked birth certificate and tumor registry data from the New York State Health Department. Cases were 319 women aged 26–45 who were diagnosed with breast cancer in New York State between 1989 and 1995 and who completed a first pregnancy in New York State after 1987 at least one year prior to diagnosis of cancer. Controls were 768 primiparous women matched to cases on county of residence and delivery date. Information on prenatal smoking and other factors characterizing the woman's first pregnancy was obtained from the pregnancy record of each subject, and the association of these factors to breast cancer risk was assessed using conditional logistic regression.
Results: Smoking during pregnancy was associated with increased risk for breast cancer (crude OR = 2.7, 95% confidence interval (CI): 1.1–6.3). Adjustment for maternal age, subject age, race, and education strengthened this association (OR = 4.8, CI 1.6–14.6).
Conclusions: These findings suggest that cigarette smoking during a woman's first pregnancy may increase her risk for early-onset breast cancer.

Introduction

Tobacco smoke promotes the formation of free-radicals [1] and has been strongly linked to the development of several human cancers [2]. However, the association of cigarette smoking with human breast cancer risk remains uncertain. While some recent studies suggest that smoking increases breast cancer risk [3–10], at least in certain subgroups [11–17], others have found no association [5, 18–23]. A few studies have suggested that smoking may even exert a weak protective influence [24, 25], perhaps due to the antiestrogenic effect of tobacco smoke [24].

Some of the variability among studies may be due to differences in age at diagnosis or to the timing of exposure relative to a woman's reproductive life cycle. Specifically, the association of smoking to breast cancer risk may be dependent at least in part on timing of exposure relative to key milestones in breast development. Smoking may be especially important during pregnancy, a period during which hormonal factors promoting both growth and differentiation of mammary tissue are dramatically elevated [26], and which may therefore be a critical period in the development of breast cancer [27]. In the non-pregnant state, the carcinogenic effects of smoking may be countered or even superseded by the antiestrogenic effects of tobacco. In contrast, during pregnancy, a time of breast hyperplasia, the increased free radical formation and other carcinogenic effects of smoking may operate synergistically with elevated levels of estrogens and other growth factors to increase breast cancer risk. These effects may be particularly important during the first pregnancy when breast tissue is less differentiated [28, 29] and hence more susceptible to mutagensis than in subsequent pregnancies [29, 30].

However, to our knowledge the relation between smoking during gestation and subsequent breast cancer

Smoking During Pregnancy and Breast Cancer

	Smokers	Non-smokers
Cases	215	1449
Controls	668	4296
Crude OR	0.95 (0.80 – 1.1)	
Adjusted OR	0.97 (0.80 – 1.2)	

BU

Could Null Result Be Due to Misclassification of Smoking on the Birth Certificate?

BOSTON UNIVERSITY

Exposure Misclassification Terms

- **Non-differential**
  - Rate of E misc don't depend on D
  - Se of E same in D+ and D-

  AND
  - Sp of E same in the D+ and D-
- **Classification values**

  Truth is the denominator, E is the truth, T is the test
  - **Sensitivity**
    - Probability of being correctly classified as E+
    - Pr(T+|E+)

- **Differential exp misclass**
  - Rates of E misc do depend on D
  - Se of E not same in D+ and D-

  OR
  - Sp of E not same in D+ and D-

- **Specificity**
  - Probability of being correctly classified as E-
  - Pr(T-|E-)

BU

Results of Validation Studies

Citation	Standard	SE	SP
Buescher	Medical record	0.87	0.99
Piper	Medical record	0.78	0.99
Dietz	Questionnaire	0.88	0.99
Dietz	Capture-recapture	0.68	1
Cancer Registry	All smokers	0.29	0.99
Cancer Registry	Current Smokers	0.47	0.99
Multiple records	Consistent Smokers	0.63	0.95

Exposure Misclassification

	Truth		Observed	
	E=1	E=0	E=1	E=0
Y=1	A	B		
Y=0				
Total				

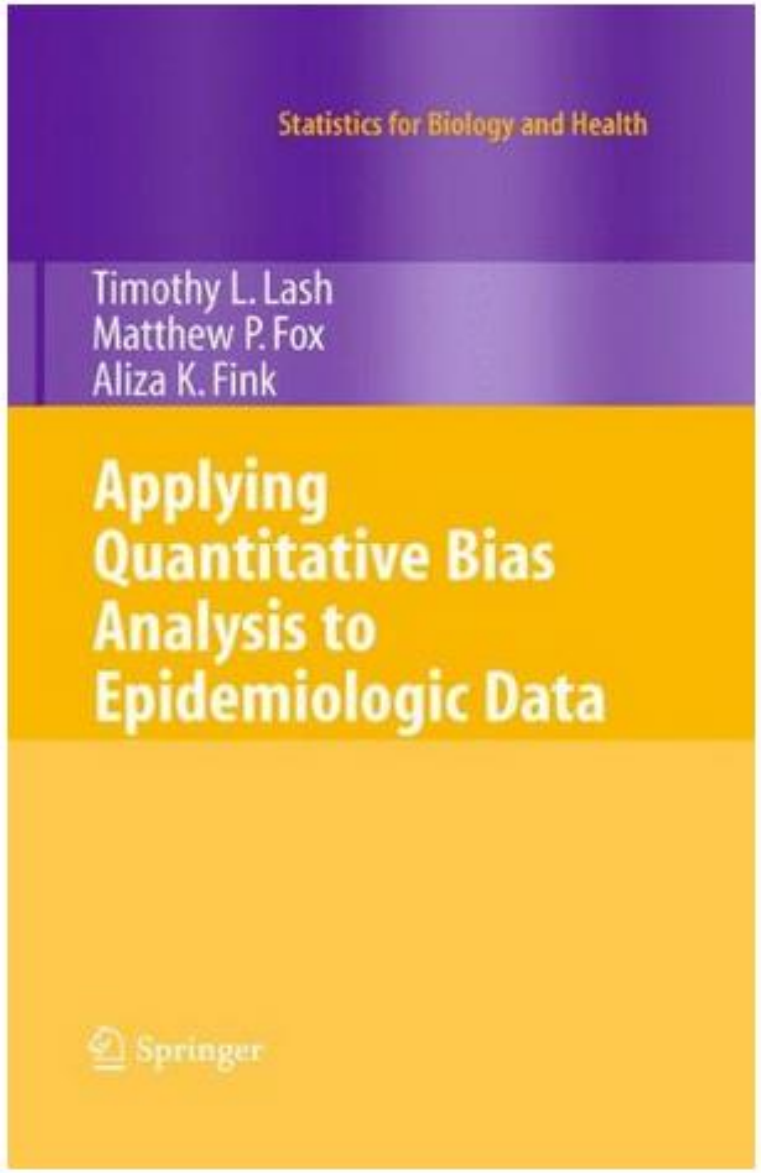
- To reverse

	Observed		Expected Truth	
	E=1	E=0	E=1	E=0
Y=1	a	b	$[a-(1-Sp_1)*Y_1] / [Se_1-(1-Sp_1))]$	Y_1 - A
Y=0	c	d	$[c-(1-Sp_0)*Y_0] / [Se_0-(1-Sp_0))]$	Y_0 - C
Total	n_1	n_0	A+C	B+D

BU

Text Book – Edition 1, 2009, Edition 2, 2022

- We will use book, but it will also be a reference for you as go back to these methods
- There are other readings, let me know if you can't access them



https://sites.google.com/site/biasanalysis/

bias.analysis

- Applying Quantitative Bias Analysis to Epidemiologic Data
  - Collaboration
  - Errata
  - Links
- Multiple Bias Model (Lash TL, Fink A)
- sensmac SAS Macro (Fox MP, Lash TL, Greenland S)

APPLYING QUANTITATIVE BIAS ANALYSIS TO EPIDEMIOLOGIC DATA

Use this page to download the accompanying spreadsheets and SAS code (see bottom of page) for the book:

Lash TL, Fox MP, Fink AK. Applying Quantitative Bias Analysis to Epidemiologic Data.

Springer. 2009.

You can find reviews of the textbook in the *American Journal of Epidemiology*, *JASA*, *JRSS* and *Biometrics*.

This book collects and synthesizes methods for quantifying systematic errors that affect observational epidemiologic research.

This text provides the first-ever compilation of bias analysis methods for use with epidemiologic data. It guides the reader through the planning stages of bias analysis, including the design of validation studies and the collection of validity data from other sources. Three chapters present methods for corrections to address selection bias, uncontrolled confounding, and classification errors. Subsequent chapters extend these

File Home Insert Page Layout Formulas Data Review View Automate Help Comments Share

T4 Calculations

EXPOSURE MISCLASSIFICATION — Chapter 6

This spreadsheet can be used to conduct a simple bias analysis to adjust for exposure misclassification. The example follows the example in chapter 6.

Reset Example | Clear Data

Input Bias Parameters

Se (BC+)	0.78
Se (BC-)	0.78
Sp (BC+)	0.99
Sp (BC-)	0.99

Corrected Variance?

Error Check: No errors

Variable Names

Outcome	BC
Exposure	Smoker

Instructions

Enter the bias parameters in the blue cells to the left and the observed data in the blue cells below. Cells in green give the results after adjusting for exposure misclassification. Note: green cells are expected values and therefore do not have to be integers.

Data (Enter Smoker-BC Data in Blue Cells)

Observed Data

	Smoker +	Smoker -	Total
BC +	215 (a)	1449 (b)	1664
BC -	668 (c)	4296 (d)	4964
Total	883 (m)	5745 (n)	

Observed	Measure (95% CI)
RR (Smoker-BC)	0.97 (0.85 - 1.09)
OR (Smoker-BC)	0.95 (0.81 - 1.13)

Bias Adjusted Data

Adjusted Data

	Smoker +	Smoker -	Total
BC +	257.6 (A)	1406.4 (B)	1664
BC -	803.1 (C)	4160.9 (D)	4964
Total	1060.7 (M)	5567.3 (N)	

Adjusted	Measure
RR (Smoker-BC)	0.96
OR (Smoker-BC)	0.95

Calculations

RR	0.965
SE(LN(RR))	0.064
OR	0.954
SE(LN(OR))	0.084
SE(LN(OR)) Adj	0.094
Adjusted OR	0.949
Pr(E*|D=1)	0.129
Pr(E*|D=0)	0.135
Errors?	0

0
0
0
0

Bounds on Bias Parameters

Se1 minimum	0.129
Se0 minimum	0.135
Sp1 minimum	0.871
Sp0 minimum	0.865

Cover | Exposure Misc | Outcome Misc | Confounder Misc | Dependent ED Misc | Dependent EC I ...

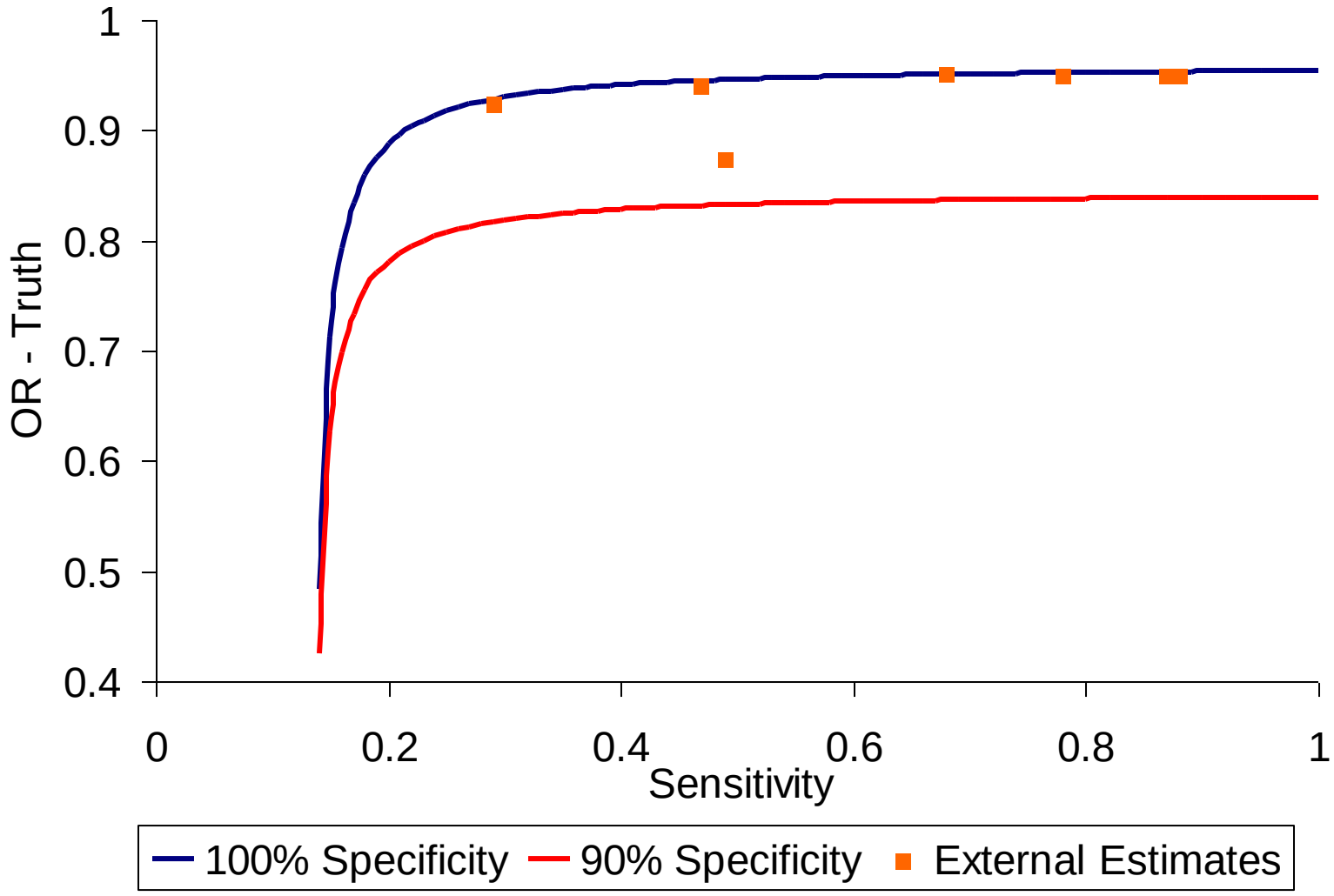
Ready Accessibility: Investigate 90%

Multidimensional Table

SE	SP	A	B	C	D	OR
0.99	1.0	217.2	1446.8	674.7	4289.2	0.95
0.50	1.0	430	1234	1336	3628	0.95
0.25	1.0	860	804	2672	2292	0.92
0.99	0.9	54.6	1609.4	192.8	4771.2	0.84
0.50	0.9	121.5	1542.5	429	4535	0.83
0.25	0.9	324	1340	1144	3820	0.82

BU

Non-differential Misclassification



Probabilistic Bias Analysis

BOSTON UNIVERSITY

Steps for Summary PBA for Misclassification*

1) Input observed, summarized data into a 2x2 table
2) Input probability distributions for sensitivity and specificity (for now, uniform)
3) Randomly sample Se / Sp from specified distributions
4) Use simple bias analysis methods previously described for exposure misclassification to adjust the observed data (fits the expectation)
5) Sample E prevalence from beta distributions parameterized by adjusted data
6) Use Se, Sp and Prevalence to calculate PPV and NPV
7) Apply PPV/NPV to sample adjusted contingency table using binomials parameterized from the adjusted data
8) Summarize the contingency table (RR, OR, RD, etc.)
9) Save the adjusted estimate
10) Repeat many times (say 50,000)
11) Add random error back in
12) Create intervals from 2.5^{th} to 97.5^{th} percentile, median as point estimate

*Note this is from edition 2 of the text – edition 1 skips steps 5-7

Orientation to Excel PBA

Ch8_Probabilistic_EDITION 2 trap_2021_03_29.xlsm

Home Insert Draw Page Layout Formulas Data Review View Tell me Share Comment

C31 =E26

PROBABILISTIC EXPOSURE MISCLASSIFICATION — Chapter 8

This spreadsheet can be used to conduct a probabilistic bias analysis to bias adjust for exposure misclassification and random error simultaneously. The example follows the example in chapter 8.

Input Bias Parameters

	Min	Mode	Mode2	Max
Se (LC+)	0.80	0.80	1.00	1.00
Se (LC-)	0.80	0.80	1.00	1.00
Sp (LC+)	0.80	0.80	1.00	1.00
Sp (LC-)	0.80	0.80	1.00	1.00
Misclassification type		ND	Sims	1000

Error: No error

Corr Se: 0.80

Corr Sp: 0.80

Instructions

Enter the bias parameter distributions in the blue cells to the left and the observed data in the blue cells below. Cells in green give the results after correcting for exposure misclassification. Note: green cells do not have to be integers.

Run Simulation

Variable Names			
Exposure	Resins	Outcome	LC

Data (Enter Resins-LC Data in Blue Cells)

Observed Data

	Resins +	Resins -	Total
LC +	45 (a)	94 (b)	139
LC -	257 (c)	945 (d)	1202
Total	302 (m)	1039 (n)	

Single Simulated Bias-adjusted Data

Bias-adjusted Data

	Resins +		Resins -		Total
LC +	43.0	A	96.0	B	139
LC -	202.0	C	1000.0	D	1202
Total	245.0	M	1096.0	N	

Chosen Values	
Se(D+)	95.7%
Se(D-)	95.7%
Sp(D+)	94.5%
Sp(D-)	94.5%

Observed	Measure (95% CI)
RR (Resins-LC)	1.65 (1.18 - 2.29)
OR (Resins-LC)	1.76 (1.2 - 2.58)

Single Bias-adjusted Estimate	Measure
RR (Resins-LC)	2
OR (Resins-LC)	2.22

RR Simulation Results (N=1000)

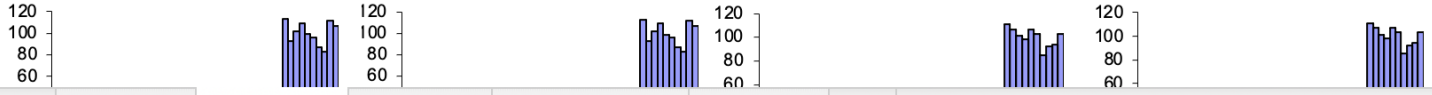
Analysis	Median (2.5th-97.5th percentile)
Conventional	1.65 (1.18 - 2.29)
Systematic	2.14 (1.69 - 5.06)
Total Error	2.15 (1.33 - 5.61)

OR Simulation Results (N=1000)

Analysis	Median (2.5th-97.5th percentile)
Conventional	1.76 (1.2 - 2.58)
Systematic	2.41 (1.81 - 8.28)
Total Error	2.47 (1.31 - 9.49)

Illogical Values
0

Chosen Se(D-) Chosen Se(D+) Chosen Sp(D-) Chosen Sp(D+)



Calculations

RR(adjusted)	2.00374	OR(adjusted)	2.217
SE(LN(RR))	0.16908	SE(LN(OR))	0.194
SE(LN(RR Adjusted))	0.16935	SE(LN(OR Adjusted))	0.199
RR rand adj	2.849	OR rand	2.621
Negative or 0 cell	FALSE		

Correlation

z1	0.78275	z1	0.596
z2	0.25708	z2	-0.271
u1	0.78311	u1	0.724
u2	0.60144	u2	0.393

Bounds on Bias Parameters

Se1 minimum	0.32374	Sp1 minimum	0.676
Se0 minimum	0.21381	Sp0 minimum	0.786

Correlations

Actual corr Se	1	Actual corr Sp	1

Single Iteration

Syst Error		+ Rand Error		Chosen Bias Parameters			
RR	OR	RR	OR	Se(D+)	Se(D-)	Sp(D+)	Sp(D-)
1.825	1.9867	2.849	2.62097	95.7%	95.7%	94.5%	94.5%
2.142	2.4066	2.388	2.12299	0.8773	0.877	0.8953	0.895
1.812	1.9622	2.379	2.76922	0.8116	0.812	0.969	0.969
2.309	2.6429	2.176	2.43027	0.8799	0.88	0.8773	0.877
1.839	2.0013	2.401	2.7575	0.8734	0.873	0.9508	0.951
1.958	2.1528	2.149	2.12213	0.8201	0.82	0.9304	0.93
1.781	1.9237	1.783	2.17698	0.836	0.836	0.9745	0.974
1.75	1.8872	1.518	1.33591	0.9	0.9	0.9742	0.974
4.684	7.334	4.314	5.02863	0.9372	0.937	0.8073	0.807
2.379	2.744	1.978	2.69263	0.8756	0.876	0.8715	0.872
1.767	1.9097	1.459	1.82823	0.9041	0.904	0.9677	0.968
2.015	2.2406	2.605	3.00855	0.9944	0.994	0.905	0.905
1.709	1.8332	2.011	1.40131	0.8545	0.855	0.9981	0.998
2.036	2.2606	3.097	2.84184	0.8505	0.851	0.9125	0.913
2.065	2.3076	1.891	3.47135	0.9712	0.971	0.8992	0.899
1.846	2.0053	2.983	2.43984	0.806	0.806	0.9598	0.96

Cover | Exp Misc | Dis Misc | Unmeas Conf | Sel Bias | +

120%

Bias Analysis Results (Col S-Y)

Syst Error		+ Rand Error		Chosen Bias Parameters			
RR	OR	RR	OR	Se(D+)	Se(D-)	Sp(D+)	Sp(D-)
2.402	2.7671	1.656	2.8814	80.5%	80.5%	87.4%	87.4%
1.877	2.0519	1.628	1.7243	0.8915	0.892	0.9391	0.939
4.34	6.4373	4.255	7.7113	0.9168	0.917	0.811	0.811
2.095	2.3481	1.874	2.8315	0.9615	0.961	0.8958	0.896
1.895	2.0718	2.152	2.8943	0.8463	0.846	0.9407	0.941
2.876	3.4965	2.827	3.602	0.8197	0.82	0.8452	0.845
1.851	2.0183	1.637	1.8645	0.9127	0.913	0.943	0.943
4.645	7.1258	4.238	5.6023	0.8464	0.846	0.8087	0.809
2.491	2.9157	2.687	2.8913	0.9191	0.919	0.8614	0.861
1.907	2.0895	1.536	2.4818	0.8794	0.879	0.9339	0.934
3.901	5.4164	2.863	7.0287	0.8771	0.877	0.8174	0.817
2.165	2.4402	2.419	2.2263	0.8927	0.893	0.8914	0.891
2.025	2.2493	2.071	2.5637	0.9131	0.913	0.9089	0.909
2.304	2.642	2.701	2.241	0.9334	0.933	0.875	0.875
1.842	2.0052	1.443	1.607	0.89	0.89	0.948	0.948
2.192	2.4861	2.734	1.9723	0.9877	0.988	0.8832	0.883
2.443	2.8359	1.966	3.1339	0.8626	0.863	0.8673	0.867
1.785	1.9343	1.888	2.7605	0.9484	0.948	0.9567	0.957

BU

Interpretation

- **Look at change in point estimate vs. conventional**
  - Direction, Magnitude
- **Focus on width of the interval**
  - Note this is not a confidence interval
  - For now focus only on systematic error
  - This is a distribution of adjusted estimates given the bias parameters

OR Simulation Results (N=1000)	
Analysis	**Median (2.5^{th}-97.5^{th} percentile)**
Conventional	1.76 (1.2 - 2.58)
Systematic	2.41 (1.81 - 8.28)
Total Error	2.47 (1.31 - 9.49)

- **Interpret interval**
  - Take the ratio of the upper and lower limit as a measure of width
  - Compare to conventional confidence interval

BU

Conclusions

- **Simple bias analysis allows is to quantify the impact of sources of bias rather than just speculating on the bias**
  - Focus on direction, magnitude
- **Probabilistic bias analysis is an improvement over simple/multidimensional bias analysis because it allows us to put realistic distributions to the bias parameters and summarize the bias with a corrected point estimate and simulation interval**
  - Focus on direction, magnitude and uncertainty

BU