Stratified Analysis:
Introduction to Confounding and Interaction

Introduction

Error in Etiologic Research

Let us briefly reconsider, in a general way, the relative risk estimates studied previously. These statistical estimates are used to estimate underlying relative risk parameters. (Recall that parameters represent error-free constants that quantify the true relationship between the exposure and disease being studied.) Unfortunately, parameters are seldom known, so we are left with imperfect estimates by which to infer them. As a general matter of understanding, we might view each statistical estimate as the value of the parameter plus "fudge factors" for random error and systematic error:

estimate = parameter + random error + systematic error

For example, a calculated relative risk of 3 might be an overestimate by 1 of the relative risk parameter, with error attributed equally to random and systematic sources: 3 (estimate) = 2 (parameter) + 0.5 (random error) + 0.5 (systematic error). Of course, the value of the parameter and its deviations-from-true are difficult (impossible) to know factually, but we would still like to get a handle on these unknowns to better understand the parameter being studied.

So what, then, is the nature of the random error and systematic error of which we speak? Briefly, *random error* can be thought of as variability due to sampling and / or the sum total of other unobservable balanced deviations. The key to working with random error is understanding how it balances out in the long run, and this is dealt with using routine methods of inference, including estimation and hypothesis testing theory.

Systematic error, in contrast to random error, is less easily managed, and less easily understood. According to one scheme, analytic systematic error (or *bias*, as they say), can be classified as either: (a) information bias, (b) selection bias, or (c) confounding.

Information bias is due to the mis-measurement or misclassification of study factors — either the exposure, the disease, or an other relevant factor. As the old saying goes, the quality of the study is limited by the quality of the measurements ("garbage in, garbage out"). Briefly, information bias can be either differential (occurring at different rates in the groups being compared) or non-differential. In general, the later is preferred, because any resulting bias due to non-differential misclassification will bring (bias) things toward the null.

Selection bias occurs as a result of nonrepresentative samples, often resulting from the use of convenience samples and other non-probability methods. The use of nonrepresentative controls in a case-control study, for example, results in a selection bias.

Confounding (from the Latin *confundere*: to mix together) is a distortion of an association between an exposure (E) and disease (D) brought about by extraneous factors (C_1, C_2, etc). This problem occurs when E is associated with C *and* C is an independent risk factor for D. For example, smoking (C) confounds the relationship between alcohol consumption (E) and lung cancer (D), since alcohol and smoking are related, and smoking (C) is an independent risk factor for lung cancer (D).

Along with confounding, we might also discuss interaction. *Interaction,* as distinct from confounding, is the interdependent operation of two or more factors to produce an unanticipated effect. We should consider statistical interaction and biological interaction separately. *Statistical interaction* occurs when a *statistical model* does not explain the joint effect of two or more independent variables. For example, if the relative risk for D associated with factor $E_1 = 2$ and the relative risk associated with factor $E_2 = 3$, we would expect under the multiplicative model suggest by relative risk for a person who has both risk factors (E_1 and E_2) to have a relative risk of 6. If the joint effects of E_1 and E_2 result in a RR other than 6, the multiplicative risk model fails to predict the association, so a statistical interaction is said to exist. Thus, interaction is model specific.

A numerical example may serve to further illuminate. Suppose that Group 0 (the unexposed group) has an average risk of 2 per 100, Group 1 (exposed to factor E_1) has an average risk of 4 per 100, and Group 2 (exposed to factor E_2) has an average risk of 6 per 100. Therefore, $RR_1 = 4 / 2 = 2$ and $RR_2 = 6 / 2 = 3$. Now suppose that persons exposed to both E_1 and E_2 have a risk of 12 per 100, so $RR_{1 \text{ and } 2} = 12 / 2 = 6$. Then, the individual relative risk are accurate in predicting joint effects, since $RR_1 \times RR_2 = 2 \times 3 = 6$, and no interaction is said to exist. Note however, that had we been working on an additive scale, using risk difference as our measure of association, then $RD_1 = 3$ per 100 - 2 per 100 = 1 per 100 and $RD_2 = 4$ per 100 - 2 per 100 = 2 per 100. The predicted combined effect would be $RD_{1 \text{ and } 2} = RD_1 + RD_2 = 1$ per 100 + 2 per 100 = 3 per 100. However, we note that $RD_{1 \text{ and } 2} = 6$ per 100 - 2 per 100 = 4 per 100. Therefore, the risk difference model (which is additive) failed to predict the joint effects of E_1 and E_2 and a statistical interaction would be said to exist on this scale. This shows how the risk difference (additive) model would show an interaction, while the same data modeled using relative risk (multiplicative risk) would show no interaction. Scenarios in the other manner (i.e., no additive interaction, but multiplicative interaction) could also be developed.

In contrast to statistical interaction, *biological interaction* occurs when there is a difference in the biologic effect of an exposure according to the presence or absence of another factor. Biological interaction can be thought of as *effect modification,* and is an example of *antagonism* and *synergy.* An example of interaction is seen in the case of oral contraceptive use (*E*), cardiovascular disease (*D*), and smoking (*C*). Because smoking (*C*) amplifies thromboembolic-disease risk (*D*) in oral contraceptive users, interaction is said to exist. This is why oral contraceptives carry a boxed warning advising against their use in smokers.

Simpson's Paradox

The idea behind Simpson's paradox is relatively simple. The investigator disaggregates the data into homogenous subgroups ("strata") to see if the association seen in the undivided, aggregate data holds true during subsequent analysis. Not surprisingly, data can apparently show one thing when they are in aggregate form and show something quite different when they are disaggregated. This phenomenon is known as Simpson's Paradox.

Measures of association in the aggregate are called *crude measures of association,* since relationships have yet to be separated out or otherwise adjusted. Let us precede acronyms with a "*c*" when referring to crude measures of association. For example, let *cRR* represent the crude relative risk (i.e., the relative risk based on an aggregate, single 2x2 table). Let us use a subscript to denote *strata-specific measures of association.* For example, let RR_1 represent the relative risk in stratum 1, RR_2 represents the relative risk in stratum 2, and so on.

Numerical illustrations will serve to demonstrate Simpson's paradox. Assume data come from a cohort study in which the exposed group shows an incidence of 200 / 1000 = 20% and the unexposed group shows an incidence of 50 / 1000 = 5%. The crude (unstratified) relative risk is therefore 20% / 5% = 4.0. However, crude relative risk may hide different patterns of risk once disaggregated. We will present three possible disaggregation scenarious consistent with this aggregate data.

Scenario A (below) shows a situation in which neither confounding nor interaction are present. Notice that the strata-specific relative risks and crude results equal 4 ($RR_1 = RR_2 = cRR = 4.0$). The need for stratification is therefore superfluous.

Stratum 1 (C+)

	D+	D-	
E+	160	240	400
E-	40	360	400

Stratum 2 (C-)

	D+	D-	
E+	40	560	600
E-	10	590	600

Pooled

	D+	D-	
E+	200	800	1000
E-	50	950	1000

RR_1 = (160 / 400) / (40 / 400) = 4.0 RR_2 = (40 / 600) / (10 / 600) = 4.0 *cRR* = (200 / 1000) / (50 / 1000) = 4.0

Scenario B shows a situation where the same crude data disaggregates to reveal strata-specific relative risks of 1.0. This suggests that the crude relative risk was confounded. Nevertheless, a single relative risk summarizes the relationship between the exposure and disease: in this case it would be safe to say *RR* = 1.

Stratum 1 (C+)

	D+	D-	
E+	194	606	800
E-	24	76	100

Stratum 2 (C-)

	D+	D-	
E+	6	194	200
E-	26	874	900

Pooled

	D+	D-	
E+	200	800	1000
E-	50	950	1000

RR_1 = (194/800) / (24/100) = 1.0 RR_2 = (6 / 200) / (26 / 900) = 1.0 *cRR* = (200/1000) / (50 / 1000) = 4.0

Now consider **Scenario C.** Notice that $RR_1 = 1.0$ and $RR_2 = 23.5$. Since the nature of the association depends on the influence of extraneous factor C, an interaction between E and C can be said to exist. In such instances, summary measures of association should be avoided in favor of the strata-specific findings.

Stratum 1 (C+)

	D+	D-	
E+	12	188	200
E-	48	752	800

Stratum 2 (C-)

	D+	D-	
E+	188	612	800
E-	2	198	200

Pooled

	D+	D-	
E+	200	800	1000
E-	50	950	1000

RR_1 = (12 / 200) / (48/800) = 1.0 RR_2 = (188/800) / (2/ 200) = 23.5 *cRR* = (200/1000) / (50 / 1000) = 4.0

Data in these scenarios illustrate how stratification might reveal otherwise hidden confounding and interaction. In fact, when we look at it this way, Simpson's paradox is not really a paradox at all, but is the logical consequence of failing to recognize the effects of an extraneous factor (Rothman, 1975) .

Illustrative Example (SEXBIAS)

To illustrate some of the concepts in this chapter, let us consider a data set collected as part of a University of California at Berkeley study to assess whether men were being given preferential treatment over women in admission to graduate programs (Bickel & O'Connell, 1975, Freedman et al., 1991, pp. 16 - 19). Assuming that the men and women who applied for admission to the graduate programs were equally well-qualified, one would expect equal acceptance rates by gender. However, it initially appeared as if men were being admitted in greater proportions than women. The experience of applicants to the six largest majors at the school is contained in SEXBIAS.REC. These data contains 4526 records with the following variables:

Variable	Description
SEX	1 = Male 2 = Female (The "exposure")
ACCEPT	Accepted into the major: +/- (The outcome)
MAJOR	Department A, B, C, D, E, or F (UC Berkeley policy does not allow majors to be identified by name)

Crude (2x2) analysis shows the following (*annotated*) results:

```
                ACCEPT
SEX             +      -    | Total
----------+---------------+------
         1 |   1198   1493 |  2691    Acceptance rate, men = 1198 /2691 = 0.445
         2 |    557   1278 |  1835    Acceptance rate, women = 557 / 1835 = 0.304
-----------+--------------+------     RR = 0.445 / 0.304 = 1.46
     Total |   1755   2771 |  4526    p < 0.00001
```

Therefore, on crude analysis, men appear to have a higher acceptance rate than women -- presumptive evidence of preferential treatment. However, we want to determine whether the more favorable acceptance rate may be due to factors other than gender. For example, what if men had applied to majors with more favorable acceptance rates. Then the MAJOR would confound the relationship between SEX and ACCEPT. To investigate this possibility, we could stratify the data by MAJOR, allowing us to look for evidence of interaction and confounding directly.

> **EpiInfo command:** TABLES <EXPOSURE> <OUTCOME> STRATAVAR= <CONFOUNDER>. For the illustrative example, the command TABLES SEX ACCEPT STRATAVAR= MAJOR is used to produce separate tables for each of the 6 majors. *Annotated* results follow:

```
                MAJOR =A
                 ACCEPT
SEX        |     +     - | Total
---------------------------------
         1 |   512   313 |   825     Acceptance rate, men = 512 / 825 = .621
         2 |    89    19 |   108     Acceptance rate, women = 89 / 108 = .824
-----------+-------------+------     RR = .621 / .824 = 0.75
     Total |   601   332 |   933     p = .000033

               MAJOR =B
                 ACCEPT
SEX        |     +     - | Total
-----------+-------------+------
         1 |   353   207 |   560     Acceptance rate, men = 353 / 560 = .630
         2 |    17     8 |    25     Acceptance rate, women =  17 /  25 = .680
-----------+-------------+------     RR = .630 / .680 = 0.93
     Total |   370   215 |   585     p = .61

               MAJOR =C
                  ACCEPT
SEX        |     +     - | Total
-----------+-------------+------
         1 |   120   205 |   325     Acceptance rate, men = 120 / 325 = .369
         2 |   202   391 |   593     Acceptance rate, women = 202 / 593 = .341
-----------+-------------+------     RR = .369 / .341 = 1.08
     Total |   322   596 |   918     p = .39

              MAJOR =D
                  ACCEPT
SEX        |     +     - | Total
-----------+-------------+------
         1 |   138   279 |   417     Acceptance rate, men = 138 / 417 = .331
         2 |   131   244 |   375     Acceptance rate, women = 131 / 375 = .349
-----------+-------------+------     RR = .331 / .349 = 0.95
     Total |   269   523 |   792     p = .59

                MAJOR =E
SEX        |     +     - | Total
-----------+-------------+------
         1 |    53   138 |   191     Acceptance rate, men = 53 / 191 = .277
         2 |    94   299 |   393     Acceptance rate, women = 94 / 393 = .239
-----------+-------------+------     RR = .277 / .239 = 1.16
     Total |   147   437 |   584     p = .32

                MAJOR =F
SEX        |     +     - | Total
---------------------------------
         1 |    22   351 |   373     Acceptance rate, men = 22 / 373 = .059
         2 |    24   317 |   341     Acceptance rate, women = 24 / 341 = .070
-----------+-------------+------     RR = .059 / .070 = 0.84
     Total |    46   668 |   714     p = .54
```

Therefore, only Major A demonstrates a significant difference in acceptance rates (in favor of women). Notice that the initial 2x2 (crude) analysis hid this pattern (Simpson's Paradox). It is now evident that MAJOR confounds the relationship between SEX and ACCEPT and interaction between SEX and MAJOR exists.

Mantel-Haenszel Methods

In the SEXBIAS.REC illustrative example we probably do not want to report one measures of association, since this would hide the association that was evident within Major A. However, in situations where confounding is present but interaction is absent, we usually want to report a single measure of association while controlling for confounding. There are several methods to perform this type of adjustment, one of which is the Mantel-Haenszel method (1959).

Let us use subscripts to denote each strata-specific table. The following standard table setup and notation is adopted:

	Disease+	**Disease-**	
Exposure +	a_i	b_i	n_{1i}
Exposure -	c_i	d_i	n_{0i}
	m_{1i}	m_{0i}	n_i

for strata i: 1 to s.

The formula for Mantel-Haenszel adjusted relative risk is:

$$\hat{RR}_{MH} = \frac{\sum_{i=1}^{s} \frac{a_i n_{0i}}{n_i}}{\sum_{i=1}^{s} \frac{c_i n_{1i}}{n_i}}$$

This provides a weighted average of the stratum-specific relative risks (without logarithmic transformation) with weights equal to $c_i n_{1i}/n_i$ and is a good approximation to the maximum likelihood estimate.

Recall, that in the illustration of Simpson's Paradox, we start with the following 2x2 table:

Pooled

	D+	**D-**	
E+	200	800	1000
E-	50	950	1000
	250	1750	2000

Therefore, the crude RR estimate = (200 / 1000) / (50 / 1000) = 4.0. However, upon stratification under illustrative scenario B, we find:

Stratum 1 (C+)

	D+	D-	
E+	194	606	800
E-	24	76	100
	218	682	900

$RR_1 = (194 / 800)/(24 / 100) = 1.0$

Stratum 2 (C-)

	D+	D-	
E+	6	194	200
E-	26	874	900
	32	1068	1100

$RR_2 = (6 / 200) / (26 / 900) = 1.0$

Therefore, the crude relative risk appears to be an artifact of confounding, and some sort of adjustment seems necessary. This adjustment is provided by the Mantel-Haenszel estimate which pools the strata-specific estimates to come up with $\hat{RR}_{MH} = \dfrac{\frac{(194)(100)}{900} + \frac{(6)(900)}{1100}}{\frac{(24)(800)}{900} + \frac{(26)(200)}{1100}} = 1.0.$

Mantel-Haenszel statistics are automatically computed when stratified tables are requested and printed toward the bottom of the output ("Summary Information" Section).

A short-cut sometime used to assess for potential confounding is to compare the crude RR to the adjusted RR, with confounding confirmed when the crude RR and adjusted RR estimates differ. Note that there is no statistical test for confounding, since confounding is a form of systematic error, not random error. Therefore, the degree to which confounding is present must remain a matter of judgment.

Confidence interval for the Mantel-Haenszel RR estimate are based on the standard error:

$$se_{\ln \hat{RR}_{MH}} = \sqrt{\frac{\sum_{i=1}^{s} m_{1i} n_{1i} n_{0i} - a_i c_i n_i / n_i^2}{\left[\sum_{i=1}^{s} \frac{a_i n_{0i}}{n_i}\right]\left[\sum_{i=1}^{s} \frac{c_i n_{1i}}{n_i}\right]}}$$

A 95% confidence interval for RR_{MH} is given by:

$$\hat{RR}_{MH}^{(\pm 1.96 \cdot se_{\ln \hat{RR}_{MH}})}$$

And a test of significance can be performed with the Mantel-Haenszel chi-square statistic, which is:

$$c^2_{MH} = \frac{\sum_{i=1}^{S} \frac{a_i d_i - b_i c_i}{n_i}}{\sum_{i=1}^{S} \frac{n_{1i} n_{0i} m_{1i} m_{0i}}{(n_i - 1) n_i^2}}$$

This chi-square statistic has 1 degree of freedom.

Overall Strategy

Although the detection and control of confounding is crucially important in etiologic research, there exists no single way for dealing with these issues.Nevertheless, several important principal exist.

First, potential confounders must be identified *before* data are collected, so that these variables can be evaluated during analysis. An essential analytic task in making decisions is to understand how things work. and an understanding of the system being investigated. Such understandings must necessarily come from previous research, clinical insight, and understanding of the disease process itself.

Second, adjustments for confounding are contraindicated when interaction is present, as such adjustments will obscure the interaction. Interactions is usually addressed by reporting data by subgroups.

Third, since confounding is a systematic (not random) error, hypothesis testing cannot be used to detect it. Determination of the presence of confounding demands an understanding how things work (mechanisms, trade-offs, processes and dynamics, cause and effect). It is a judgement based science.

And a few additional points:

a. One should attempt to understand the complex inter-relationships among all the determinants of the disease being studied. This may require close collaboration among subject-matter specialist.

b. Study designs and measurements that maximize the validity of the study are essential; are first and foremost.

c. After data are collected, entered and cleaned, the analyst should start with simple comparisons of means and proportions. An understanding of the relationships among the multiple factors will heighten the awareness and potential for confounding.

d. Stratified analyses are a fundamental element of most causal thinking. Eplorations for interaction are among the first applied. When interaction is confirmed, strata-specific estimates are reported.

e. Confounding is considered by controlling for extraneous factors and determining the “effect” of such controls.

f. In the absence of interaction and confounding, crude (unadjusted) estimates of association may be reported.

g. The best estimate of association is both valid and precise. If interaction is present, strata-specific measures of association are reported. If interaction is absent but confounding is present, summary (adjusted) measures of association are reported. If neither interaction nor confounding are present, crude (unadjusted) measures of association are reported. *In general, the most parsimoniously unconfounded presentation of the data is preferred.* If the association between the exposure and disease is not found by scrutinizing the data in the 2-by-2 table, it's hard to support. Simple is better.

Study Questions

1. Define *confounding*.

   ANS: Confounding is a distortion of an association between an exposure and disease brought about by extraneous factors.

2. Define *interaction*.

   ANS: Interaction is the interdependent operation of two or more factors to produce an unanticipated effect.

3. What preconditions are necessary for confounding?

   ANS: The preconditions for confounding are: (a) *E* and *C* must be associated, *and* (b) *D* and *C* must be associated.

4. How does one check for interaction when using stratified tables to measure the association between an exposure and disease?

   ANS: One approach is to check for interaction by stratifying the data on the potentially interactive factor. Strata-specific measures of association (i.e., strata-specific risk ratios or odds ratios) are then compared. Significant differences in strata-specific measures of association imply that interaction is present. A chi-square test for interaction may be used to help confirm this presence.

5. How does one check for confounding?

   ANS: One checks for confounding only after interaction has been ruled out. Confounding can be assessed in several different ways. The most direct way is to carry through an adjustment technique (e.g., by calculating the Mantel-Haenszel summary measure of effect) and compare this adjusted estimate to the crude results (e.g., unadjusted measure of effect produced by simple 2-by-2 table analysis). If these estimates are similar, confounding is probably *absent*. If, on the other hand, these estimates differ, confounding is present. Note that there is no formal hypothesis test upon which to base the presence of confounding. The decision as to the presence of confounding is based upon whether or not the adjustment changes one's interpretation of the data.

6. How does one report results when interaction is present?

   ANS: When interaction is present, report separate results for strata defined by the presence and absence of the effect modifier.

7. How does one report results when confounding is present?

   ANS: When confounding is present, report adjusted *RR*s (cohort studies) and adjusted *OR*s (case-control studies).

8. How does one report results when neither interaction nor confounding are present?

   ANS: Report crude (2-by-2 table) results.

9. Why do we report strata-specific measures of association when interaction is present?

ANS: Because there is no overall measure of association (associations vary from subgroup to subgroup).

10. Why do we report adjusted measures of association when confounding is present?

ANS: Because the crude measure of association is biased ("confounded").

11. Why do we report only crude estimates when neither interaction nor confounding are present?

ANS: Because this provides the most precise, ubiased measure of association.

12. What is the purpose of the Mantel-Haenszel procedure?

ANS: The Mantel-Haenszel procedure provides a summary measure of association by adjusting for confounders when interaction is absent.

References

Bickel, P. & O'Connell, J. W. (1975). Is there a sex bias in graduate admissions? *Science*, 187, 398 - 404.

Breslow, N. E., & Day, N. E.(1980). *Statistical Methods in Cancer Research. Volume 1--The Analysis of Case-Control Studies.* Lyon: International Agency for Research on Cancer.

Freedman, D., Pisani, R., Purves, R., & Adhikari, A. (1991). *Statistics* (2nd ed.) New York: W. W. Norton.

Gerstman, B. B., Jolson, H., Bauer, M., Cho, P., Livingston, J., & Platt R. (1996). Depression in new users of ß-blockers and selected anti-hypertensives. *Journal of Clinical Epidemiology*, 49, 809 - 815.

Hirayama, T. (1990). *Life-style and Mortality: a Large Scale Census-based Cohort Study in Japan*. Basel: S. Karger.

Kneale, G. W. (1971). Problems arising in estimating from retrospective survey data the latent period of juvenile cancers initiated by obstetric radiography. *Biometrics*, 27, 563 - 90.

Kramer, M. S. (1988). *Clinical Epidemiology and Biostatistics.* Berlin: Springer-Verlag.

Lilienfeld, D. E. & Stolley, P. D. (1994). *Foundations of Epidemiology* (3rd ed.). New York: Oxford.

Mandel, E., Bluestone, C. D., Rockette, H. E., Blatter, M. M., Reisinger, K. S., Wucher, F. P., & Harper, J. (1982). Duration of effusion after antibiotic treatment for acute otitis media: comparison of cefaclor and amoxicillin. *Pediatric Infectious Diseases*, 1, 310 - 316.

Mantel, N., Haenszel, W.. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, 22, 719 - 748.

Nishan, P., Ebeling, K, Schindler C. (1988). Smoking and invasive cervical cancer risk: results from a case-control study. *American Journal of Epidemiology*, 128, 74 - 77.

Pagano, M. & Gauvreau, K. (1993). *Principles of Biostatistics.* Belmont, CA: Duxbury Press.

Perales, D. & Gerstman, B. B. (1995, March). A bi-county comparative study of bicycle helmet knowledge and use by California elementary school children. *The Ninth Annual California Conference on Childhood Injury Control*, San Diego, CA.

Rosner, B. (1990). *Fundamentals of Biostatistics* (3rd ed.) Boston: PWS - Kent Publishing.

Rothman, K. J. (1975). A pictorial representation of confounding in epidemiologic studies. *Journal of Chronic Diseases*, 28, 101 - 108.

Stewart, A. & Kneale, G. W. (1970). Age-distribution of cancers caused by obstetric X-rays and their relevance to cancer latent periods. *Lancet*, *ii*, 4 - 8.

Tuyns, A. J., Péquignot, G., & Jensen, O. M.. (1977). Le cancer de l'oesophage en Ille-et Vilaine en function des niveaux de consommation d'alcool et de tabac. Des risques qui se multiplient. *Bull Cancer*, 64, 45 - 60.