Evidence Factors For Observational Studies: Design, Analysis And Computation

Bikram Karmakar
University of Pennsylvania, bikramk@wharton.upenn.edu

Follow this and additional works at: https://repository.upenn.edu/edissertations
Part of the Epidemiology Commons, Quantitative, Qualitative, Comparative, and Historical Methodologies Commons, and the Statistics and Probability Commons

Recommended Citation
https://repository.upenn.edu/edissertations/3217

This paper is posted at ScholarlyCommons. https://repository.upenn.edu/edissertations/3217
For more information, please contact repository@pobox.upenn.edu.
Evidence Factors For Observational Studies: Design, Analysis And Computation

Abstract
This thesis includes five chapters on evidence factors analysis of causal effect in various observational study settings. Each of these chapters can be read independently without knowledge of the content of any of the other chapters. Evidence factors allow for two independent analyses to be constructed from the same data set. When combining the evidence factors, the type-I error rate must be controlled to obtain valid inference. A powerful method is developed for controlling the familywise error rate for sensitivity analyses to unmeasured confounding with evidence factors. It is shown that the Bahadur efficiency of sensitivity analysis for the combined evidence is greater than for either evidence factor alone. The popular strategy of matching, for controlling the observed covariates, before inferring about the treatment effect, requires solving an optimization problem. This problem can be solved in polynomial time. In an evidence factors analysis we must consider multiple comparisons, thus the matching problem is often of matching at least three groups. This slightly different problem is much more difficult to solve. The third chapter proposes an approximation algorithm to solve this (and more practical versions of this) problem. We prove that the proposed algorithm provides a solution fast, that is provably not a lot further than the optimal solution that is difficult calculate. Two chapters that follow show the applicability of evidence factors analysis in more complicated study designs. The first of these two chapters considers a case-control study with multiple case definitions and the latter one considers studies with instrumental variables, where the instrument(s) may become invalid. The final chapter of the thesis develops a frequentist method for quantification of the degree of corroboration of causal hypothesis using the tool of evidence factors.

Degree Type
Dissertation

Degree Name
Doctor of Philosophy (PhD)

This dissertation is available at ScholarlyCommons: https://repository.upenn.edu/edissertations/3217
Graduate Group
Statistics

First Advisor
Dylan S. Small

Keywords
Causal inference, Evidence factors, Instrumental variables, Observational study, Sensitivity analysis to unmeasured confounding, Treatment effect

Subject Categories
Epidemiology | Quantitative, Qualitative, Comparative, and Historical Methodologies | Statistics and Probability

This dissertation is available at ScholarlyCommons: https://repository.upenn.edu/edissertations/3217
EVIDENCE FACTORS FOR OBSERVATIONAL STUDIES: DESIGN, ANALYSIS AND COMPUTATION

Bikram Karmakar

A DISSERTATION

in

Statistics

For the Graduate Group in Managerial Science and Applied Economics

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2019

Supervisor of Dissertation

Dylan S. Small, Class of 1965 Wharton Professor of Statistics

Graduate Group Chairperson

Catherine Schrand, Celia Z. Moh Professor, Professor of Accounting

Dissertation Committee

Paul R. Rosenbaum, Robert G. Putzel Professor of Statistics

Jason Roy, Chair & Professor Department of Biostatistics and Epidemiology, Rutger University
ACKNOWLEDGEMENT

I would like to thank my adviser and mentor Professor Dylan Small for his enormous support in my work throughout my doctoral study. I would not have learned so much of this fascinating topic of causal inference if I had not taken his class in the second semester of my first year as a graduate student at Penn. He has been extremely helpful in developing my research interest. More than that, I have learned so much life lesson from him. One of life goals is to one day be as dedicated in pursuing my passion as I have seen Dylan to be.

One of my biggest achievements during my years at Wharton has been to meet and work with Professor Paul Rosenbaum. Through our hours long meetings every week I have learned more than I could ever learn about causal inference. He is the most intelligent person I have met in my life. The way Paul, Dylan and I would work on a problem through discussions and debates has have been the highlights of my week, each week.

I want to thank each of my committee members, Dylan Small, Paul Rosenbaum and Jason Roy for their support in my work. When I am working on my research, or, more likely, annoyed by not being able make any progress, each of them has been there to help me.

I must thank Maggie Saia and Gidget Murray of Wharton Doctoral Programs who have cheerfully guided me through the administrative process of obtaining my degree. I cannot but thank my cohort-mates, Gemma, Raiden, Justin and Linjun. I wish them all the success that I know for certain they will achieve in their lives.

I have gotten to know so many incredible people at Penn. I thank each of them for their generosity with their time and for bearing with my stupidity. I must again include Dylan and Paul in this list. My thanks in particular to Mark Low, Ed George, Larry Brown, and Andreas Buja. I feel spoiled by the help and support of the wonderful stuff of the Statistics Department, particularly, Carol, Noelle, Tanya and Sarin.

Last but not the least, I owe all of whatever little I have been able to make of myself this
far to my family, my mother, my father and my brother. My parents’ have been by my side for most of my life. Now, even though I live thousands of miles away from them, their teachings are the compass of my life.

When I first jointed Wharton as a young graduate student, someone had told me something that has stuck with me. I paraphrase it here: “You think it’s gonna be easy? It’s not. You’ll need every bit of support you can get to make it through.” I might not have paid too much attention to these words then, but no other sets words can be more true in a graduate students life, in my opinion. Finally, my extremely apologies if I have missed someone.
ABSTRACT

EVIDENCE FACTORS FOR OBSERVATIONAL STUDIES: DESIGN, ANALYSIS AND COMPUTATION

Bikram Karmakar
Dylan S. Small

This thesis includes five chapters on evidence factors analysis of causal effect in various observational study settings. Each of these chapters can be read independently without knowledge of the content of any of the other chapters. Evidence factors allow for two independent analyses to be constructed from the same data set. When combining the evidence factors, the type-I error rate must be controlled to obtain valid inference. A powerful method is developed for controlling the familywise error rate for sensitivity analyses to unmeasured confounding with evidence factors. It is shown that the Bahadur efficiency of sensitivity analysis for the combined evidence is greater than for either evidence factor alone. The popular strategy of matching, for controlling the observed covariates, before inferring about the treatment effect, requires solving an optimization problem. This problem can be solved in polynomial time. In an evidence factors analysis we must consider multiple comparisons, thus the matching problem is often of matching at least three groups. This slightly different problem is much more difficult to solve. The third chapter proposes an approximation algorithm to solve this (and more practical versions of this) problem. We prove that the proposed algorithm provides a solution fast, that is provably not a lot further than the optimal solution that is difficult calculate. Two chapters that follow show the applicability of evidence factors analysis in more complicated study designs. The first of these two chapters considers a case-control study with multiple case definitions and the latter one considers studies with instrumental variables, where the instrument(s) may become invalid. The final chapter of the thesis develops a frequentist method for quantification of the degree of corroboration of causal hypothesis using the tool of evidence factors.
# TABLE OF CONTENTS

ACKNOWLEDGEMENT ................................................................. iii

ABSTRACT .................................................................................. v

LIST OF TABLES ........................................................................... xii

LIST OF ILLUSTRATIONS ............................................................... xiv

CHAPTER 1 : Introduction .............................................................. 1

CHAPTER 2 : Integrating the evidence from evidence factors in observational studies 4

  2.1 Introduction ........................................................................ 4
  2.2 Example: solid cancer incidence in atomic bomb survivors ........ 5
  2.3 Evidence factors: a general viewpoint ................................. 8
  2.4 Combining evidence .......................................................... 11
  2.5 Integrating evidence ......................................................... 12
  2.6 Simulation: combined evidence does better in finite sample .... 19
  2.7 Analysis of the Life Span Study data ................................. 22
  2.8 Discussion ........................................................................ 25

CHAPTER 3 : Using Approximation Algorithms to Build Evidence Factors and Related Designs for Observational Studies 26

  3.1 The need for approximation algorithms when constructing an observational design .............................................. 26
  3.2 Motivating Application: Effects of Side Airbags in Crashes ...... 28
  3.3 Problem, algorithm, and guarantee ..................................... 36
  3.4 Covariate distances that satisfy the needed triangle inequality . 45
  3.5 Distinguishing effects of side airbags and unmeasured biases .... 47
3.6 Discussion .......................................................... 51

CHAPTER 4: Evidence Factors in a Case-Control Study with Application to the
Effect of Flexible Sigmoidoscopy Screening on Colorectal Cancer . 53
4.1 Introduction ......................................................... 53
4.2 Sigmoidoscopy and colorectal cancer ........................... 58
4.3 Notation and review: case-control studies ...................... 61
4.4 Two case definitions and two comparisons ..................... 63
4.5 Evidence factors ................................................. 67
4.6 Evidence factors with differential effect of unmeasured confounders on the
factors ................................................................. 73
4.7 Results: efficacy of screening sigmoidoscopy ................... 77
4.8 Discussion ........................................................ 79

CHAPTER 5: Reinforced designs for observational studies of treatment effects:
Multiple instruments plus control groups as evidence factors ................ 81
5.1 Introduction: active, effective concern for bias in observational studies 81
5.2 How effective are Catholic schools compared to public school? ........ 84
5.3 Notation and background: several instruments plus a direct comparison .. 86
5.4 Evidence factors: Combining the K steps ....................... 93
5.5 Evaluating the performance of the proposed analysis ............... 96
5.6 Effects of Catholic versus public high schools ................... 102
5.7 Discussion ........................................................ 106

CHAPTER 6: Assessment of the extent of corroboration of an elaborate theory of
a causal hypothesis using partial conjunctions of evidence factors . 108
6.1 Introduction ......................................................... 108
6.2 Lead absorption study of Morton et al. ........................... 112
6.3 Matched pair design with multiple treatments across pairs ............ 114
6.4 Evidence factors and pooling evidence .......................... 120
6.5 Evidence from a partial conjunction of the tests: A quantification of the extent of corroboration .................................................. 122
6.6 Comparison of combining methods ........................................ 128
6.7 Revisiting the lead absorption study ..................................... 135
6.8 Conclusion ............................................................................. 137

APPENDIX A .................................................................................. 139
APPENDIX B .................................................................................. 149
APPENDIX C .................................................................................. 151
BIBLIOGRAPHY ............................................................................. 157
TABLE 1: Solid cancer incidences among atomic bomb survivors in Hiroshima

TABLE 2: Sensitivity analysis for the hypothesis of no carcinogenic effect of radiation versus harmful effect of radiation

TABLE 3: Balance on covariates in 1398 matched 1-1-1 matched triples and 978 matched 1-3-1 sets. Cars were also matched for make and model. Group “none” refers to an era when side airbags were not available for this make/model, “optional” to an era when side airbags were optional for this make/model, and “all” to an era when all cars of this make/model had side airbags. For driver’s age, crash year, and model year, values are means; otherwise, they are percentages. A mean is computed within a matched set, then averaged over sets.

TABLE 4: Balance on the covariates in the matched sets. Distal cancer cases are those who have been diagnosed to have died from cancer on the left colon or rectum, proximal cancer cases are from right colon cancer. For each covariate the mean is calculated within a matched set, then averaged over sets.

TABLE 5: Simulated power of a sensitivity analysis of combined evidence in a case control study, where there is no unmeasured confounder and \( Pr(Z_l = 1) = 1/3 \). The response is simulated from \( N(\beta, 1) \) if \( Z_l = 1 \) and \( N(0, 1) \) if \( Z_l = 0 \). There are 1,000 narrow cases and 1,000 marginal cases with 2,000 controls. Based on 10,000 iterations. Fisher = Fisher’s combination method, tP = truncated product method with \( \tilde{\alpha} = 0.20 \)
TABLE 6: Simulated power of a sensitivity analysis of combined evidence in a case control study, where there is no unmeasured confounder and \( Pr(Z_l = 1) = 1/3 \). The response is simulated from \( \beta + t_3/\sqrt{3} \) if \( Z_l = 1 \) and \( t_3/\sqrt{3} \) if \( Z_l = 0 \). There are 1,000 narrow cases and 1,000 marginal cases with 2,000 controls. Based on 10,000 iterations. Fisher = Fisher’s combination method, tP = truncated product method with \( \tilde{\alpha} = 0.20 \).

TABLE 7: Screening sigmoidoscopy and colorectal cancer summary data.

TABLE 8: Counts and percents attending Catholic school for two potential instruments and a direct comparison of Catholic and public schools.

TABLE 9: Simulated probability of rejecting, at the 0.05 level, the true null hypothesis \( H_0 : \beta = \beta_0 \) using two-stage least squares (TSLS), (i) using the first binary instrument alone, (ii) using the second binary instrument stratifying for the first instrument, (iii) using a direct comparison of treated and control groups stratifying for both instruments. Because \( H_0 \) is true, every rejection is a mistake. If the test is valid (V), then the probability of rejection is 0.05, but if the test is biased (B) then the probability of rejection is inflated. The evidence-factor analysis fails to provide a warning if all three factors, (i)-(iii), are biased (B), which occurs in three cases, 8, 15 and 16. TSLS fails in 12 cases.

TABLE 10: Design sensitivities \( \tilde{\Gamma} \) for valid analyses. For biased analyses, a “B” is given in place of a design sensitivity.
TABLE 11: Three evidence factors and their combination using the truncated product, with and without covariance adjustment. The case \( \Gamma = 1 \) assumes comparisons are flawless, three stratified randomized experiments. The table shows one equivalent amplification of each \( \Gamma > 1 \). The table displays upper bounds on one-sided \( P \)-values testing the null hypothesis that Catholic schooling raises wages by at most \( \beta \) dollars in the presence of a bias of at most \( \Gamma \). As the median annual wage was $14000, a $500 increase is about 3.6%.

TABLE 12: Simulation results for power of sensitivity analysis evaluated at level 0.05. Numbers are out of 100. A cell value is the percentage of times the decision that at least \( k \) many \( H_{0,i} \)'s are false is made, with \( \Gamma_1 = \cdots = \Gamma_5 =: \Gamma \), out of 15000 simulations. Empty cells represent the value 0. tP = truncated product method with truncation level \( z = 0.20 \); sP = the modified additive p-value method in Proposition 6; Si = Simes’ method; HB = Holm-Bonferroni method.

TABLE 13: The p-values, under the assumption of no unmeasured confounding, for testing the hypothesis that at least \( k \) many \( H_{0,i} \)'s are false in the lead absorption study. \( K = 5 \) and \( \Gamma_1 = \cdots = \Gamma_5 = 1 \). SumP = the modified additive p-value method in Proposition 6.

TABLE 14: Evidence factors analysis of the lead absorption study. (1) The first half of the table: Maximum p-values corresponding to the five tests with \( \Gamma_k = \Gamma, \ 1 \leq k \leq 5 \). We dropped the subscript \( a_{k-1} \) from \( P_{k,a_{k-1},\Gamma} \) used in Section 3.3–Section 6. (2) The second half of the table: Maximum p-values for testing at least \( k \) of \( H_{0,i} \)'s are false when the bias is at most \( \Gamma_1 = \cdots = \Gamma_5 = \Gamma \) (using the truncated product method; truncation level \( z = 0.20 \)). The maximum p-values less than 0.05 are highlighted in bold.
TABLE 15 : Power of sensitivity analysis for evidence factors for Fisher’s combination method and the truncated product method based on \( S = 200 \) strata. Estimation of power is based on 1000 iterations. \( \Gamma = \Gamma_1 = \Gamma_2.140 \)

TABLE 16 : Power of sensitivity analysis for evidence factors for Fisher’s combination method and the truncated product method based on \( S = 500 \) strata. Estimation of power is based on 1000 iterations. \( \Gamma = \Gamma_1 = \Gamma_2.141 \)
LIST OF ILLUSTRATIONS

FIGURE 1: Illustration of the collection of the hypotheses in a finite sample. Gray area is the null hypotheses and the hatched area depicts the hypotheses not rejected based on the data. ..................... 14

FIGURE 2: Graphs for combined evidence (solid), high to low dose (dashes) and exposed versus unexposed (dots). Panel (A) plots the negative of the rate of rejection of the hypothesis, $-\log(\text{probability of rejection})/S$, in the null scenario, against the number of strata ($S$). Panels (B) and (C) plot the negative of the rate of acceptance of the hypothesis, $-\log(\text{probability of acceptance})/S$ against $S$. Along the rows $\Gamma_1$ and $\Gamma_2$ are varied. Results are based on average over 2000 simulations and over a grid of $S$ values in gaps of 20. . 21

FIGURE 3: Result for testing radiation effect on solid cancer. In decreasing order of gradient the colors represent the decision - reject for both comparisons, for high to low dose comparison, for city to not-in-city comparison, without any attribution and do not reject. Bonferroni method rejects the null if $\Gamma_1 < 1\cdot16$ and $\Gamma_2 < 1\cdot11$ (dashed lines). . 23

FIGURE 4: Driver’s characteristics before matching and balance of the characteristics after matching across the three eras, None, Optional, and All. Bars of similar height after matching indicate that matching has balanced the covariate. . . . . . . . . . . . . . . . . . . 34

FIGURE 5: Characteristics of the crash before matching and balance of the characteristics after matching across the three eras, None, Optional, and All. Bars of similar height after matching indicate that matching has balanced the covariate. . . . . . . . . . . . . . . . . . . 35
FIGURE 6: Injury severity in matched crashes grouped by the three eras, None, Optional, and All. The star in the boxplot and the vertical line in the barplot represent the mean.

FIGURE 7: Unmatched crashes in the optional era grouped by whether side airbags were present. The star in the boxplot and the vertical line in the barplot represent the mean. This interesting but unmatched comparison is not an evidence factor.

FIGURE 8: Reference date of the colorectal cancer cases and controls in the matched sets.

FIGURE 9: Level of bias from unmeasured confounding plotted under three speculations – bias only from $v_1$, plotted on the x-axis and in ‘red’; bias only from $v_2$, plotted on the y-axis and in ‘blue’; and biases from both $v_1$ and $v_2$, plotted in ‘green’ contours. The contours are of the function $f(\delta_{v_1}, \delta_{v_2}) = (1/\delta_{v_1} + 1/\delta_{v_2})^{-1}$. Here, $\psi_1 = 3/4$, $\psi_2 = 1/4$, $\tilde{\psi}_1 = 1/5$ and $\tilde{\psi}_2 = 4/5$. The bias levels $\delta_{bc}$ and $\delta_{nm}$ changes with the speculation and the required bias level is minimized when biases from both $v_1$ and $v_2$ are assumed.

FIGURE 10: Sensitivity analysis of the efficacy of screening sigmoidoscopy in reducing mortality from colorectal cancer. The darker gray color represents the bias levels where the combined evidence for a beneficial effect of screening sigmoidoscopy is sensitive.

FIGURE 11: Correspondence of the sensitivity parameter $\Gamma$ with the pair of parameters $(\Delta, \Lambda)$. 
CHAPTER 1: Introduction

The most well recognized model for causal inference for a treatment effect is the potential outcomes model of Neyman (1923) and Rubin (1974). Under this model, the treatment effect of a unit is the difference of the potential outcome of the unit under treatment and the potential outcome when the unit is unexposed to the treatment. For a given unit, only one of these two potential outcomes can be observed, based on whether the unit is treated or remains unexposed. Thus, the fundamental problem of causal inference is the fact that both potential outcomes of any unit is never observed simultaneously. Consequently, the treatment effect of the unit cannot be observed. The solution to this problem is the achieved through “randomization”. For example, for a pair of units, if the treatment is assigned with a flip of a fair coin to one of the units while the other remains in control, then over this randomization the average observed outcomes of the treated unit compared to the control unit estimates the average treatment effect of these two units.

The context of this thesis is causal inference in observational studies. In an observational study, the treatment assignment is not randomized. Hence, a randomization inference most likely will provide a biased estimate of the average treatment effect. The randomization may be more plausible after adjusting pretreatment covariates. But, after adjusting for these observed pretreatment covariates, say by matching, stratification, or parametric adjustments, there is still potential for unmeasured confounders that can bias the inference. In this context, this thesis develops a tool for strengthening evidence for a causal effect from an observational study. This tool called evidence factors. It is motivated by the idea of replicability of a scientific study.

A scientific evidence is repeated if the same study is conducted in a difference situation or time, and the scientific evidence from the two studies agree with each other. A replicated evidence is stronger than a repeated evidence. A replicated evidence for a fact under investigation is evidence from multiple studies investigating the fact with independently
but also using different scientific methods. This replicated evidence is stronger primarily because it consists of independent pieces of evidence that depend on very different study structures. One of these studies might be challenged, as a scientific study often is, but it would not invalidate the replicated evidence. A study of a causal effect is a study of a scientific fact under investigation. An evidence factors analysis for a causal effect forms internally replicated pieces of evidence from one observational study.

This thesis includes five chapters on evidence factors analysis of causal effect in various observational study settings. Each of these Chapters can be read independently without knowledge of the content of any of the other chapters.

Chapter 2 provides a general introduction to evidence factors analysis. This chapter is published in Biometrika, 2019, doi:10.1093/biomet/asz003. It focuses on combining the pieces of evidence from the evidence factors. In this process the type-I error rate must be controlled to obtain a valid inference. A powerful method is developed for controlling the familywise error rate for sensitivity analyses with evidence factors. It is shown that the Bahadur efficiency of sensitivity analysis for the combined evidence is greater than for either evidence factor alone. The proposed methods are illustrated through a study of the effect of radiation exposure on risk of cancer using the data from the Radiation Effects Research Foundation. This chapter is a joint work with Prof. Ben French and Prof. Small. The Radiation Effects Research Foundation (RERF), Hiroshima and Nagasaki, Japan is a public interest foundation funded by the Japanese Ministry of Health, Labour and Welfare (MHLW) and the U.S. Department of Energy (DOE). The research was also funded in part through DOE award DE-HS0000031 to the National Academy of Sciences. This publication was supported by RERF Research Protocol RP-S3-16. The views of the authors do not necessarily reflect those of the two governments.

Chapter 3 focuses on the design part of an evidence factors analysis, in an observational study. The popular strategy of matching for controlling the observed covariates before inferring about the treatment effect requires solving an optimization problem. This problem
can be solved easily in polynomial time. In an evidence factors analysis we must consider multiple comparisons, thus the matching problem is often of matching at least three groups. In theory, this slightly different problem (from the two group matching problem) is a much more difficult problem to solve. This chapter proposes an approximation algorithm to solve this (and more practical versions of this) problem. We prove that the proposed algorithm provides a solution fast that is provably not a lot further than the optimal solution that is difficult calculate in general. This chapter is sourced from an upcoming article in the *Journal of Computational and Graphical Statistics*, a joint work with Prof. Dylan Small and Prof. Paul Rosenbaum.

Chapter 4 and 5 present the applicability of evidence factors analysis in more complicated study designs. The first of these two chapters considers a case-control study with multiple case definitions and the latter one considers studies with instrumental variables, where the instrument(s) may become invalid. These chapters are based on work done in collaboration with Prof. Small, Prof. Rosenbaum and Prof. Chyke Doubeni respectively. The work of Chapter 4 was supported by an award (number R01CA213645 and number U01CA151736) from the National Cancer Institute of the National Institute of Health. The views expressed there are those of the authors only and do not represent any official position of the National Cancer Institute or National Institutes of Health.

The final chapter, Chapter 6, of the thesis considers the question of testing for a causal hypothesis, which is the broader epistemological question of causality. The positivistic philosophy of knowledge generation suggests that the causal hypothesis should be studied through investigation of several its falsifiable basic statements. How is the evidence for the causal hypothesis assessed? There is not one unambiguously accepted answer to this question. In the literature, the answer has survived most critic is called the ‘degree of corroboration’ ([Popper, 1959](#)). This final chapter of the thesis develops a frequentist method for quantification of the degree of corroboration of causal hypothesis using the tool of evidence factors.
CHAPTER 2: Integrating the evidence from evidence factors in observational studies

2.1. Introduction

In an observational study, treatment assignment is typically assumed to be effectively random conditional on measured covariates. However, the presence of unmeasured confounding can result in non-random treatment assignment, such that standard analysis methods can provide biased estimates of treatment effects. The potential for measured and unmeasured confounding motivates consideration of sensitivity analyses to assess how much bias, due to non-random treatment assignment, would be necessary to change the conclusions of a randomization inference (Cornfield et al., 1959; Rosenbaum, 1987; Keele and Minozzi, 2013; Stuart et al., 2013; Ding & VanderWeele, 2016; McCandless and Gustafson, 2017).

Evidence factors – two or more independent tests that could be sensitive to different biases – provide an approach to strengthen the evidence for a treatment effect (Rosenbaum, 2010a, 2011). When considering sensitivity analyses with evidence factors, multiple comparisons arise by performing more than one test of the same null hypothesis and by considering different sensitivity parameters. Therefore, multiplicity error must be controlled to obtain valid inference. Previous research has not considered the impact of multiplicity error when generating inference based on evidence factors. Standard methods for multiplicity adjustment, such as a Bonferroni correction, could impose a harsh penalty when there are multiple sources of evidence. Consideration of evidence factors is meant to strengthen the aggregate evidence for a treatment effect, but a punitive penalty for multiple comparisons can hamper the ability to detect a significant treatment effect. Can the attractive benefits of evidence factors analysis always be obtained? This chapter provides an affirmative answer. We provide a powerful and computationally fast method for combining evidence factors that controls for multiplicity. We show that, in terms of the Bahadur efficiency of the sensitivity analysis, our approach for combining evidence from multiple sources has better performance than considering any of the sources separately.
2.2. Example: solid cancer incidence in atomic bomb survivors

Understanding the health effects of radiation exposure is important for establishing recommendations for radiation protection, including limits on occupational exposure to radiation and guidelines for diagnostic and therapeutic use of radiation. Because randomized experiments on humans are unethical, observational studies are a key resource for estimating radiation effects.

In 1945, the United States detonated two atomic bombs over the Japanese cities of Hiroshima and Nagasaki. The Life Span Study investigates the long-term health effects of radiation exposure among survivors of the atomic bombings. The Life Span Study includes: proximal survivors, who were within 3 km of the hypocenter; distal survivors, who were between 3 and 10 km of the hypocenter; and city residents who were not in either city at the time of the bombings, and therefore not exposed to radiation. A survivor’s radiation dose is estimated from a dosimetry system that accounts for the survivor’s reported location and shielding at the time of the bombing, with the total dose given by the sum of the $\gamma$-ray dose and 10 times the neutron dose in units of gray (Gy) \cite{Cullings2017}.

Following Preston et al. \cite{Preston2007}, our goal is to evaluate the hypothesis that radiation increases the risk of solid cancer. At the time of the bombings, there were notable differences between Hiroshima and Nagasaki. Hiroshima, with even terrain, was an embarkation port and a site of major military headquarters, whereas Nagasaki, with varied terrain, was a center of heavy industry, with associated air pollution. To minimize heterogeneity, our analysis was limited to those Life Span Study participants from Hiroshima alive and at risk for solid cancer as of January 1, 1958, when population-based cancer registries were established. In addition, we did not consider distal survivors because of concern that distal survivors, who lived in more rural areas, and proximal survivors, who lived in more urban areas, could have different cancer rates for reasons other than radiation dose \cite{PiercePreston2000}.

To assess the effect of radiation exposure on the risk of solid cancer, one could compare
proximal survivors with high doses to proximal survivors with low doses; the validity of this comparison relies on the assumption that proximal survivors with low and high doses are similar on all characteristics other than their radiation exposure. This assumption could be violated because the hypocenter was close to the urban center, so that the proximal survivors with high doses tended to be located in more urban areas; also, high-dose survivors might have been comparatively healthier to have survived a high dose \cite{Preston et al., 2007}. Alternatively, one could compare cancer rates between proximal survivors and not-in-city residents; the validity of this comparison relies on the assumption that proximal survivors and not-in-city residents are similar on all characteristics other than their radiation exposure. This assumption could be violated, for example, if not-in-city residents were better educated or employed. Although both of these comparisons use the proximal survivors, we will show that under the null hypothesis of no effect, they are nearly independent in the sense that their p-values are stochastically as large as the p-values from two independent comparisons under the null hypothesis, which are uniform on the unit square. This will be discussed more formally in §2.3.1 and additional details provided in Appendix A.

Table 1 provides a summary of these two comparisons. The incidence rate of solid cancer was 18.0% among proximal survivors exposed to low doses versus 35.3% among those exposed to high doses, amounting to an 18.1% incidence rate among all proximal survivors. Among not-in-city residents, the incidence rate was 16.4%. After matching on age and sex, 58 strata were created for the low-dose versus high-dose comparison among proximal survivors and 30 strata were created for the comparison of all proximal survivors versus not-in-city residents. Both the comparisons give strong evidence suggesting radiation exposure is harmful, with
Table 2: Sensitivity analysis for the hypothesis of no carcinogenic effect of radiation versus harmful effect of radiation

<table>
<thead>
<tr>
<th>Γ</th>
<th>Maximum p-value</th>
<th>Γ₁</th>
<th>Γ₂</th>
<th>joint evidence</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>High-dose vs. Low-dose survivors</td>
<td>1</td>
<td>1</td>
<td>1.47×10⁻¹¹</td>
</tr>
<tr>
<td>1</td>
<td>0.0021</td>
<td>1</td>
<td>1</td>
<td>0.00032</td>
</tr>
<tr>
<td>1.2</td>
<td>0.0443</td>
<td>1</td>
<td>1.2</td>
<td>0.01420</td>
</tr>
<tr>
<td>1.3</td>
<td>0.1166</td>
<td>1.2</td>
<td>1</td>
<td>2.73×10⁻¹⁰</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1.2</td>
<td>1</td>
<td>0.00491</td>
</tr>
<tr>
<td></td>
<td>Proximal survivors vs. NIC residents</td>
<td>1.2</td>
<td>1.2</td>
<td>0.17117</td>
</tr>
<tr>
<td>1</td>
<td>2.35×10⁻¹⁰</td>
<td>1.3</td>
<td>1</td>
<td>6.94×10⁻¹⁰</td>
</tr>
<tr>
<td>1.1</td>
<td>0.0131</td>
<td>1.3</td>
<td>1.1</td>
<td>0.01145</td>
</tr>
<tr>
<td>1.2</td>
<td>0.9207</td>
<td>1.3</td>
<td>1.2</td>
<td>0.34688</td>
</tr>
</tbody>
</table>

NIC, not-in-city.

one-sided p-values from Mantel–Haenszel tests 0.0021 and 2.35×10⁻¹⁰ respectively.

What is the gain from considering two p-values from two analyses? Each p-value is computed based on an assumption of no unmeasured confounders for the given comparison. This assumption could be violated for one comparison but not the other. For example, there might be unmeasured differences between people who lived near the hypocenter of the bomb, high-dose proximal survivors, versus far, low-dose proximal survivors, but not between people who were in or out of the city at the time of the bombing, or vice versa. If both p-values indicate strong evidence against the null hypothesis, then there would have to be unmeasured confounders for both comparisons in order to bring the results into question.

For each of the two comparisons, we associate a single sensitivity parameter measuring the bias due to the presence of unmeasured confounders. One can study the effect of potential bias by evaluating the strength of evidence for a treatment effect from the comparison for different values of the sensitivity parameter. This sensitivity parameter is defined as the maximum odds that, among two participants with the same measured confounders, one participant would receive treatment and the other control compared to vice versa because of differences in unmeasured confounders (e.g. Rosenbaum, 1987). In §2.3.2 a formal definition of this sensitivity parameter is given. When the value of this sensitivity parameter is 1, it indicates the assumption of no unmeasured confounders; a value of 2 would mean...
that the unmeasured confounders can double the odds of receiving treatment. Let $\Gamma_1$ and $\Gamma_2$ denote these sensitivity parameters for the two comparisons. Table 2 reports the maximum $p$-values for the two comparisons for different values of these parameters. When $\Gamma_1 = 1.2$ and $\Gamma_2 = 1.1$, both the comparisons reject the null hypothesis with maximum $p$-values 0.0443 and 0.0131 respectively. The evidences from the two comparisons are sensitive at bias levels $\Gamma_1 = 1.3$ and $\Gamma_2 = 1.2$, respectively. Table 2 also reports the joint evidence given $(\Gamma_1, \Gamma_2)$ values. The joint evidence is calculated using Fisher’s combination method. If $(\Gamma_1, \Gamma_2) = (1.2, 1.2)$ we fail to reject the null hypothesis. The existing theory of evidence factors only allows us to make these statements about a given pair $(\Gamma_1, \Gamma_2)$. An objective statistician would not choose a value of $(\Gamma_1, \Gamma_2)$, but rather present the results for a range of values, in particular focusing on the value where the inference is sensitive. Hence, we would like to make a comprehensive statement for a range of values of $(\Gamma_1, \Gamma_2)$ while ensuring that the familywise error rate is controlled.

2.3. Evidence factors: a general viewpoint

2.3.1. Definition of evidence factors

Suppose we wish to test a hypothesis $H_0$ and let $A_1$ and $A_2$ be two different assumptions under which the hypothesis can be tested. Let the evidence gathered against $H_0$ based on $A_1$ be $E_1$ and the evidence based on $A_2$ be $E_2$ after taking out $E_1$. If $E_1$ and $E_2$ are $p$-values calculated from the data given the assumptions $A_1$ and $A_2$ respectively, then $E_1$ and $E_2$ would constitute separate evidence factors if they are independent upon the assumption of $A_1 \cap A_2$. Henceforth in our discussion by evidence against null we mean the (maximum) $p$-value. The requirement of independence can be relaxed because the desired property of $(E_1, E_2)$ is: when considered jointly they provide more evidence against $H_0$ than separately. The pair $(E_1, E_2)$ are called evidence factors if, when both $A_1$, $A_2$ and $H_0$ hold, the joint cumulative distribution function of $(E_1, E_2)$ is stochastically larger than the joint distribution of two independent $p$-values. As shown in §2.3.2, this definition implies that, most tests for the null hypothesis using the evidence factors can use the cutoff calculated
assuming independence and be valid. Since \( p \)-values are uniformly distributed under the null hypothesis, this amounts to having for all \((p_1, p_2) \in [0, 1]^2\)

\[
\Pr(E_1 \leq p_1, E_2 \leq p_2) \leq p_1 \times p_2.
\]  

(2.1)

**Definition 1** A set \( D \subseteq \mathbb{R}^k \) is called a decreasing set if for any \( x, y \in \mathbb{R}^k \) with \( x \leq y \), if \( y \in D \) then \( x \in D \). For two random vectors \( X \) and \( Y \) we say that \( X \) is stochastically larger than \( Y \), in notation \( X \succ Y \), if \( \Pr(X \in D) \leq \Pr(Y \in D) \) for all decreasing sets \( D \).

**Definition 2** The pair \((E_1, E_2)\) is said to form evidence factors for testing \( H_0 \) assuming \( A_1 \) and \( A_2 \) if, \((E_1, E_2) \succ (U_1, U_2)\) under \( A_1 \cap A_2 \) and \( H_0 \), for two independent \( \text{Unif}[0,1] \) random variables \( U_1 \) and \( U_2 \).

Since \([0, p_1] \times [0, p_2]\) are decreasing sets, if \((E_1, E_2)\) are evidence factors then (2.1) is satisfied.

2.3.2. **Sensitivity analysis and evidence factors**

Consider the sensitivity of the evidences \( E_1 \) and \( E_2 \) with respect to their corresponding assumptions \( A_1 \) and \( A_2 \). Let \( \Gamma_1(\geq 1) \) be a real number that quantifies possible deviation from assumption \( A_1 \) (Gastwirth, 1992; Hosman et al., 2010; Zubizarreta et al., 2012; Rosenbaum, 2002, §4). For instance, when the assumption \( A_1 \) is that the treatment is randomly assigned among the treated and the control units, i.e. there are no unmeasured confounders, \( \Gamma_1 \) would quantify bias in treatment assignment due to possible unmeasured confounders. To make this precise, we discuss one definition of the parameter \( \Gamma_1 \) (Rosenbaum, 2002, §4) here. Let \( i = 1, \ldots, n \) be the indices assigned arbitrarily to \( n \) units. Let \( Z_i \) be the indicator for unit \( i \) being in the treatment group. Also, let \( x_i \) denote the observed pretreatment covariates while \( u_i \) is an unobserved number summarizing the unobserved confounders for unit \( i \) (see Rosenbaum, 1987). Finally, suppose unit \( i \) if exposed to treatment would have response \( r_{T_i} \) and if spared exposure would have response \( r_{C_i} \).
Consequently, both $r_{Ti}$ and $r_{Ci}$ are not observed simultaneously (Neyman, 1923). Let $\mathcal{F} = \{(r_{Ti}, r_{Ci}, x_i, u_i) | i = 1, \ldots, n\}$. Then the sensitivity parameter $\Gamma_1$ will be defined by,

$$
\Gamma_1 = \max_{1 \leq i, i' \leq n; x_i = x_{i'}} \Pr(Z_i = 1 | \mathcal{F}) \Pr(Z_{i'} = 0 | \mathcal{F}) \{ \Pr(Z_i = 1 | \mathcal{F}) \Pr(Z_i = 0 | \mathcal{F}) \}^{-1}.
$$

In words, the model for treatment assignment is such that due to unobserved covariates the odds of being treated for two units $i$ and $i'$ with the same observed covariates is allowed to differ at most by a multiplicative factor $\Gamma_1 (\geq 1)$. When there are no unmeasured confounders, $\Gamma_1 = 1$ and two units similar in terms of their observed covariates have the same probabilities of receiving treatment. Let $A_1(\Gamma_1)$ denote all treatment assignment distributions that deviate from this randomized assignment, assumption $A_1$, by bias level at most $\Gamma_1$.

In a more general setup, when testing $H_0$ based on a test statistic $T_1$, the set $A_1(\Gamma_1)$ would specify a family of possible distributions $\mathcal{P}_1(\Gamma_1)$ for $T_1$, and the larger $\Gamma_1$ is, the larger the family of distributions $\mathcal{P}_1(\Gamma_1)$ becomes. Define sensitivity parameter $\Gamma_2 (\geq 1)$ and corresponding $A_2(\Gamma_2)$ similarly for the second factor. Thus, the larger $\Gamma_j$ is, the more uncertain we are about the design and $\Gamma_j = 1$ implies that we are certain about the aspect $A_j$ of the design, i.e. $A_j(1) = A_j$ for $j = 1, 2$.

The sensitivity analysis computes the largest possible $p$-values under $A_1(\Gamma_1)$ and $A_2(\Gamma_2)$ as $E_1\{A_1(\Gamma_1)\}$ and $E_2\{A_2(\Gamma_2)\}$. Naturally, a larger uncertainty about the design will lead to weaker evidence. Because these assumptions are nested; for $j = 1, 2$, with $\Gamma_j \leq \Gamma'_j$

$$
A_j(\Gamma_j) \subseteq A_j(\Gamma'_j), \quad E_j\{A_j(\Gamma_j)\} \leq E_j\{A_j(\Gamma'_j)\}.
$$

Following §2.3.1, $\{E_1\{A_1(\Gamma_1)\} | \Gamma_1 \geq 1\}$ and $\{E_2\{A_2(\Gamma_2)\} | \Gamma_2 \geq 1\}$ are said to form evidence factors for testing $H_0$ if for any $\Gamma_1$ and $\Gamma_2$ the pair $(E_1\{A_1(\Gamma_1)\}, E_2\{A_2(\Gamma_2)\})$ constitute evidence factors under the assumptions $A_1(\Gamma_1)$ and $A_2(\Gamma_2)$. We use the shorthand notation $E_{1,\Gamma_1}$ and $E_{2,\Gamma_2}$ for $E_1\{A_1(\Gamma_1)\}$ and $E_2\{A_2(\Gamma_2)\}$, respectively, because there is no ambiguity.
2.4. Combining evidence

How should we quantify combined evidence against $H_0$ from the evidence factors? Fisher’s method, which is used in §2.2, is a natural choice.

**Lemma 1** Under $H_0$, the distribution of $-2 \log E_1 E_2$ is stochastically smaller than the $\chi^2$-distribution with 4 degrees of freedom.

**Proof.** Let $U_1, U_2$ be two independent Unif[0, 1] random variables. For $0 \leq p, q \leq 1$, $pq = \Pr(U_1 \leq p, U_2 \leq q)$. Further, $-2 \log U_1 U_2$ is distributed as $\chi^2$ with 4 degrees of freedom. As $(p, q) \mapsto -2 \log pq$ is a monotone function in both coordinates by Theorem 6.B.16 of [Shaked and Shanthikumar (2007), §6], $-2 \log E_1 E_2$ is stochastically smaller than $\chi^2_4$ distribution. ■

Since, by definition, $(E_{1, \Gamma_1}, E_{2, \Gamma_2})$ form evidence factors, the combined evidence for bias levels $\Gamma_1$ and $\Gamma_2$, calculated using Fisher’s method is $E_{\Gamma_1, \Gamma_2} = \Pr(\chi^2_4 > -2 \log(E_{1, \Gamma_1} E_{2, \Gamma_2}))$.

An alternative method of combining $p$-values is [Zaykin et al. (2002)]’s truncated product method, which puts more emphasis than Fisher’s method on looking for small $p$-values. The truncated product for some $\tilde{\alpha} \in (0, 1)$ is defined as

$$E_{\tilde{\alpha}, \Gamma_1, \Gamma_2} = 1_{E_{1, \Gamma_1} \leq \tilde{\alpha}} \log E_{1, \Gamma_1} + 1_{E_{2, \Gamma_2} \leq \tilde{\alpha}} \log E_{2, \Gamma_2};$$

with $1_A$ denoting the indicator of an event $A$. An evidence factor contributes to $E_{\tilde{\alpha}, \Gamma_1, \Gamma_2}$ only if the evidence from that factor is strong, i.e., less than $\tilde{\alpha}$. Fisher’s method corresponds to $\tilde{\alpha} = 1$. [Hsu et al. (2013)] presented simulations and discussion that suggested that the truncated product method often performs better than Fisher’s method in sensitivity analysis. The following lemma studies the null distribution of $E_{\tilde{\alpha}, \Gamma_1, \Gamma_2}$. 

11
Lemma 2  Let \( W \) be a random variable on \([0, \alpha^2]\) with the distribution function

\[
F_W(w) = 2\alpha(1 - \alpha)(1 - F_{Exp(1)}[-\log \{w(\alpha)^{-1}\}]) + \alpha^2(1 - F_{Gamma(2,1)}[-\log \{w(\alpha)^{-2}\}]).
\]

Then under \( H_0 \) and \( A_1(\Gamma_1) \cap A_2(\Gamma_2) \), \( \exp(E_{\Lambda,\Gamma_1,\Gamma_2}) \) is stochastically larger than \( W \).

Proof. Hsu et al. (2013) provided a simple argument to prove the lemma in the case where \( E_{1,\Gamma_1} \) and \( E_{2,\Gamma_2} \) were independent. Define \( f_\Lambda(x, y) = \exp\{x \leq \alpha \log x + 1, y \leq \alpha \log y\} \). Then, \( f_\Lambda \) is a monotone nondecreasing function. Because the pair \((E_{1,\Gamma_1}, E_{2,\Gamma_2})\) form evidence factors, by Theorem 6.B.16 of Shaked and Shanthikumar (2007, §6), \( \exp(E_{\Lambda,\Gamma_1,\Gamma_2}) \succ f_\Lambda(U_1, U_2) \).

Using this fact and the argument of Hsu et al. the proof of the lemma follows. \( \blacksquare \)

Based on the truncated product \( E_{\Lambda,\Gamma_1,\Gamma_2} \), \( H_0 \) is rejected at level \( \alpha \) if \( \exp(E_{\Lambda,\Gamma_1,\Gamma_2}) \) is smaller than the \( \alpha \)th quantile of the distribution \( F_W \). The combined evidence \( E_{\Gamma_1,\Gamma_2} \) is quantified as \( F_W(\exp E_{\Lambda,\Gamma_1,\Gamma_2}) \). The choice of \( \alpha \) is more subjective. Choosing \( \alpha = 0.10 \) and \( 0.20 \) has been advised (Hsu et al., 2013; Zaykin et al., 2002).

Other methods of combining \( p \)-values, where the combination is increasing in \( E_{j,\Gamma_j} \), can be used, e.g. the mean of normal transformations of the evidences defined as \( \Phi\{w^{1/2}\Phi^{-1}(E_{1,\Gamma_1}) + (1 - w)^{1/2}\Phi^{-1}(E_{2,\Gamma_2})\} \) (Liptak, 1958). Which method is best for combining \( p \)-values remains unsettled. Littell & Folks (1971) show that asymptotically, in terms of Bahadur efficiency, Fisher’s combination method is optimal. Won et al. (2009) and Whitlock (2005) both show that with appropriate choice of weights, Liptak’s method has more power than Fisher’s method. Becker (1994) provides a comprehensive survey of various methods for combining \( p \)-values.

2.5. Integrating evidence

2.5.1. Sensitivity analysis and familywise error rate control

A sensitivity analysis for increasing, potentially infinite, sequences of \( \{\Gamma_{1i} \mid i = 1, \ldots\} \) and \( \{\Gamma_{2i} \mid i = 1, \ldots\} \) values involves tests of multiple hypotheses. For a pair \((\Gamma_{1i}, \Gamma_{2i'})\) the
hypothesis being tested is

\[ H_{0,\Gamma_1,\Gamma_2'} : H_0 \cap A_1(\Gamma_1) \cap A_2(\Gamma_2'). \]

In words, \( H_{0,\Gamma_1,\Gamma_2'} \) is the hypothesis that \( H_0 \) is true and the deviation from assumption \( A_1 \) is at most \( \Gamma_1 \) and from \( A_2 \) is at most \( \Gamma_2' \). Since multiple hypotheses are tested simultaneously, controlling type-I error is a concern. Fortunately, as shown below, the structure of the problem allows us to perform each test at level \( \alpha \) while controlling for total error at \( \alpha \).

Let \( \bar{\Gamma}_j = \min\{\Gamma_{ji} \mid i = 1, \ldots; A_j(\Gamma_{ji}) \text{ is true}\} \) for \( j = 1, 2 \) with the convention that the minimum of an empty sequence is infinity. When \( H_0 \) is true, \( H_{0,\Gamma_1,\Gamma_2'} \) is true if and only if \( \Gamma_1 \geq \bar{\Gamma}_1 \) and \( \Gamma_2' \geq \bar{\Gamma}_2 \). In Fig. 1 the shaded gray area denotes the set of true null \( \mathcal{H}_0 = \{H_{0,\Gamma_1,\Gamma_2'} \mid \Gamma_1 \geq \bar{\Gamma}_1, \Gamma_2' \geq \bar{\Gamma}_2\} \). Let \( E_{\Gamma_1,\Gamma_2'} \) denote the combined evidence against \( H_0 \) under \( A_1(\Gamma_{1i}) \cap A_2(\Gamma_{2i'}) \). In sensitivity analysis on a single parameter, take \( \Gamma_1 \) for example, under \( H_0 \) a false rejection at level \( \Gamma_1 \geq \bar{\Gamma}_1 \) implies false rejection at \( \bar{\Gamma}_1 \) which is controlled at level \( \alpha \). Thus the familywise error rate for sensitivity analysis on a single parameter is controlled at the desired level. The following theorem shows that the same argument generalizes for more than one parameter when the parameters correspond to different evidence factors.

**Theorem 3** Suppose \( E_{\Gamma_1,\Gamma_2'} \) is a nondecreasing function of individual evidences. Consider the testing procedure where \( H_{0,\Gamma_1,\Gamma_2'} \) is rejected if and only if \( E_{\Gamma_1,\Gamma_2'} < \alpha \). Then the probability of rejecting any \( \mathcal{H}_0 \) is at most \( \alpha \).

**Proof.** If \( H_0 \) is false there is nothing to prove. Recall property (2.2) of individual evidences. Now, the joint evidence is nondecreasing in individual evidences. As a consequence of these facts the retained set of hypothesis, i.e. the set of \( H_{0,\Gamma_1,\Gamma_2'} \) with \( E_{\Gamma_1,\Gamma_2'} \geq \alpha \), must be an increasing convex set as depicted in form of gridded area in Fig. 1. Thus, for the proposed testing procedure under \( H_0 \), \( Pr(\text{any } \mathcal{H}_0 \text{ is rejected}) = Pr(H_{0,\Gamma_1,\Gamma_2'} \text{ is rejected for some } \Gamma_1 \geq \)
Figure 1: Illustration of the collection of the hypotheses in a finite sample. Gray area is the null hypotheses and the hatched area depicts the hypotheses not rejected based on the data.
\[ \Pr(H_0, \Gamma_1, \Gamma_2 \text{ is rejected}) = \Pr(E_{\Gamma_1, \Gamma_2} < \alpha) \leq \alpha. \] The first inequality follows from convexity of the retention set; the second inequality uses the fact that \( E_{\Gamma_1, \Gamma_2} > \text{Unif}[0,1] \).

Methods of combining evidence described in §4 (Fisher’s method, the truncated product method) all satisfy the condition of Theorem 3 that \( E_{\Gamma_1, \Gamma_2} \) is a nondecreasing function of the individual. However other methods, such as a modified Liptak’s method where \( w \) is a function of \( j \varphi \) as in Chen and Nadarajah (2014) do not.

The retention set of biases, \( \{ (\Gamma_{1i}, \Gamma_{2i'}) \mid E_{\Gamma_{1i}, \Gamma_{2i'}} \geq \alpha \} \), has a nice structure – it is a convex and increasing set. As a consequence, this set can be computed in \( O(\log \max_{j=1,2} |\mathcal{G}_j|) \) time, where \( |\mathcal{G}_j| \) is the range for bias on \( j \)th evidence factor. This benefit is substantial when there are \( d \) evidence factors each with a finite possible bias range \( |\mathcal{G}_j| \). Then the complexity, \( O(d \log \max_{j=1}^d |\mathcal{G}_j|) \), is linear in \( d \) as compared to \( O(\prod_{j=1}^d |\mathcal{G}_j|) \) for linear search algorithms.

The appendix of this chapter includes pseudo code for such an algorithm. We have written an R package evidenceFactors, available on CRAN, that implements this algorithm along with other methods of this chapter.

The above result is more general than stated. One can restrict attention to special subsets of the grid and still ensure multiplicity control. For example Pimentel et al. (2015) discusses testing for pairs \( \{ (\Gamma_{1i}, 0.80\Gamma_{1i}) \mid i = 1, \ldots \} \). Proof of the following corollary is given in Appendix A.

**Corollary 1** Let \( E_{\Gamma_{1i}, \Gamma_{2i'}} \) be as in Theorem 5. Let \( \mathcal{G} \) be a fixed continuous subset of \( \{ \Gamma_{1i} \mid i \geq 1 \} \times \{ \Gamma_{2i} \mid i \geq 1 \} \) such that \( \mathcal{G} \cap \{ (1, \Gamma_{2i}) \mid i \geq 1 \} \cup \{ (\Gamma_{1i}, 1) \mid i \geq 1 \} \) is non-empty. Then, the probability that the testing procedure of Theorem 5 on \( \mathcal{G} \) falsely rejects any hypothesis is at most \( \alpha \).
2.5.2. Design sensitivity, consistency and asymptotic rate

Most problems of testing have a design sensitivity attached to them, which is an asymptotic measure of power of sensitivity analysis that is not dependent on $\alpha$ (see Rosenbaum, 2004).

The design sensitivity is the level of bias above which the power goes to zero as the sample size goes to infinity for any significance level, and below which the power goes to one. The design sensitivity, denoted by $\tilde{\Gamma}$, is the value of the sensitivity parameter at which the corresponding test can asymptotically distinguish a treatment effect with no bias from any treatment effect with bias less than $\tilde{\Gamma}$, but not from no treatment effect with bias larger than $\tilde{\Gamma}$. Let $\tilde{\Gamma}_1$ and $\tilde{\Gamma}_2$ denote the design sensitivity of the first and second kind of biases respectively. Then by definition, for $j = 1, 2$; $E_{j, \Gamma_j} \to 1$ for $\Gamma_j > \tilde{\Gamma}_j$, and $E_{j, \Gamma_j} \to 0$ for $\Gamma_j < \tilde{\Gamma}_j$. All convergence statements here and later are in almost sure sense as the sample size goes to infinity. To explicitly show the dependence on sample size $n$ we write $E_{j, \Gamma_j|n}$ and $E_{\Gamma_1, \Gamma_2|n}$. A consequence of the above is that the joint evidence satisfies: $E_{\Gamma_1, \Gamma_2|n} \to 0$ if $\Gamma_1 < \tilde{\Gamma}_1$ or $\Gamma_2 < \tilde{\Gamma}_2$; $E_{\Gamma_1, \Gamma_2|n} \to 1$ if $\Gamma_j > \tilde{\Gamma}_j$ for both $j = 1, 2$. Pictorially, this means that the gridded area in Fig. 1, which is the collection of hypotheses not rejected, in the limiting case, with the sample size going to infinity, will coincide with the gray rectangular area depicting the collection of true null hypotheses ($H_0$). Hence, the design sensitivity of the joint conclusion is $(\tilde{\Gamma}_1, \tilde{\Gamma}_2)$.

However, these limits do not provide any information on the rates at which such convergences take place. One can consider the Bahadur slope (Bahadur, 1967), which is the rate of convergence of $E_{j, \Gamma_j}$ on a logarithmic scale. For example, if it exists, the Bahadur slope for $\Gamma_j < \tilde{\Gamma}_j$ would be $\lim_{n \to \infty} n^{-1} \log E_{j, \Gamma_j|n}$ for the $j$th evidence factor and for the joint evidence it would be $\lim_{n \to \infty} n^{-1} \log E_{\Gamma_1, \Gamma_2|n}$. Rosenbaum (2015c) introduced the Bahadur efficiency of sensitivity analysis in this context. Taking cue from that discussion, we consider the probability of large deviation in rejection and acceptance decisions for the evidences. As shown by Rosenbaum (2015c), existence of an exact rate depends on the test statistic used. But an upper bound of the rate can always be considered. 

16
§4.5. Let \( I_{j,\gamma_j} \) be functions defined on \([0, 1]\) taking non-negative values, including possibly \( \infty \), such that, for any compact subset \( F \) of \([0, 1]\), for \( j = 1, 2 \)

\[
\limsup_{n \to \infty} n^{-1} \log \Pr(E_{j,\gamma_j}^n \in F) \leq - \inf_{x \in F} I_{j,\gamma_j}(x).
\]  

(2.3)

Because \( \tilde{\Gamma}_j \) is the design sensitivity of \( j \)th factor, if \( \Gamma_j > \tilde{\Gamma}_j \) we would expect \( I_{j,\gamma_j}(x) > 0 \) for any \( x < 1 \) and if \( \Gamma_j < \tilde{\Gamma}_j \) we would expect \( I_{j,\gamma_j}(x) > 0 \) for any \( x > 0 \). In quantitative terms (2.3) says, when \( \Gamma_j > \tilde{\Gamma}_j \) the probability of rejecting the null based on \( j \)th factor is less than \( \varepsilon \) for sample sizes more than \( \log(1/\varepsilon)/\inf_{x \in [0,1]} I_{j,\gamma_j}(x) \). Similarly if \( \Gamma_j < \tilde{\Gamma}_j \) the probability of failing to accept the null based on evidence \( j \) is less than \( \varepsilon \) for \( n > \log(1/\varepsilon)/\inf_{x \in [0,\alpha]} I_{j,\gamma_j}(x) \).

We wish to establish that the joint test has a rate which is larger than that of the individual tests. Theorem 4 requires \((E_{1,\Gamma_1}, E_{2,\Gamma_2})\) to be evidence factors in the following sense

\[
(E_{1,\Gamma_1|n}, E_{2,\Gamma_2|n}) \succ (\tilde{E}_{1,\Gamma_1|n}, \tilde{E}_{2,\Gamma_2|n}),
\]

(2.4)

where \( \tilde{E}_{1,\Gamma_1|n} \) and \( \tilde{E}_{2,\Gamma_2|n} \) are independently distributed and \( \tilde{E}_{j,\Gamma_j|n} \) have the same distribution as \( E_{j,\Gamma_j|n} \). While Definition 2 uses stochastic ordering under \( H_0 \), (2.4) is a more general statement also under the alternative hypothesis.

**Theorem 4** Suppose \( I_{j,\gamma_j} \) satisfies (2.3) for \( j = 1, 2 \). Then with \( \alpha < 0.20 \), for Fisher’s combination

\[
\limsup_{n \to \infty} n^{-1} \log \Pr(E_{\Gamma_1,\Gamma_2}^n < \alpha) \leq - \inf_{x: x \leq \alpha} \max_{j=1,2} I_{j,\gamma_j}(x),
\]  

(2.5)

\[
\limsup_{n \to \infty} n^{-1} \log \Pr(E_{\Gamma_1,\Gamma_2}^n \geq \alpha) \leq - \inf_{x: x \geq \alpha} \max_{j=1,2} I_{j,\gamma_j}(x).
\]  

(2.6)

Since \( E_{j,\Gamma_j|n} \), for \( j = 1, 2 \), converges to zero or one almost surely, with \( \tilde{\alpha} \) fixed, Fisher’s method and truncated product are equivalent for large \( n \). Thus, Theorem 4 holds for the truncated products method as well. Theorem 4 does not assume that evidence factors are well behaved, i.e. does not assume that in (2.3) the limit of \( n^{-1} \log \Pr(E_{j,\Gamma_j|n} \in F) \) exists. It
allows us to make claims about the worst rates, e.g. in terms of \( \limsup n^{-1} \log \Pr(E_{\Gamma_1, \Gamma_2|n} \geq \alpha) \) and \( \limsup n^{-1} \log \Pr(E_{\Gamma_1, \Gamma_2|n} < \alpha) \). If in (2.3), \( \limsup \) can be replaced by \( \lim \) and equality in place of inequality, hence the exact rates of rejection and acceptance for the factors exists, then both (2.5) and (2.6) hold with \( \limsup \) replaced by \( \lim \). Theorem 4 can be interpreted as: the joint evidence requires a smaller sample size to make the correct decision than the factors considered separately. An illustration of this result is given through simulation in §2.6. The proof of Theorem 4 is given in the supplement.

If the evidence factors are well behaved, more accurate statements about the rates can be made. Theorem 5 indicates that if individual factors have Bahadur slopes, then the Bahadur slope of the joint evidence is again better than the individuals.

**Theorem 5** Suppose for a pair \( (\Gamma_1, \Gamma_2) \), there exits two non-negative numbers \( r_{1, \Gamma_1} \) and \( r_{2, \Gamma_2} \) such that: (i) \( n^{-1} \log E_{\Gamma_1, \Gamma_2|n} \to -r_{1, \Gamma_1} \), and (ii) \( n^{-1} \log E_{2, \Gamma_2|n} \to -r_{2, \Gamma_2} \). Then for Fisher’s combination method, \( \lim_{n \to \infty} n^{-1} \log E_{\Gamma_1, \Gamma_2|n} = -(r_{1, \Gamma_1} + r_{2, \Gamma_2}) \). Also, if for some non-negative \( a_{j, \Gamma_j} \), \( n^{-1} \log (1 - E_{j, \Gamma_j|n}) \to -a_{j, \Gamma_j} \), for \( j = 2 \) (or 1), then with (i) (or (ii)), \( \lim_{n \to \infty} n^{-1} \log E_{\Gamma_1, \Gamma_2|n} = -r_{j, \Gamma_j} \), where \( j = 1 \) (or 2).

**Proof.** Recall that \( E_{\Gamma_1, \Gamma_2|n} = \Pr(\chi_4^2 > -2 \log E_{\Gamma_1, \Gamma_2|n} E_{\Gamma_2|n}) \). For any \( t \), as \( n \to \infty \), \( n^{-1} \log \Pr(\chi_4^2 > nt^2) \to -t^2/2 \). Now under (i) and (ii), \( -2n^{-1} \log E_{\Gamma_1, \Gamma_2|n} E_{\Gamma_2|n} \to 2(r_{1, \Gamma_1} + r_{2, \Gamma_2}) \). Let \( c \) and \( d \) be any numbers such that \( c < (r_{1, \Gamma_1} + r_{2, \Gamma_2}) < d \), then, \( 2c < -2n^{-1} \log E_{\Gamma_1, \Gamma_2|n} E_{\Gamma_2|n} < 2d \) for large enough \( n \). Thus for large \( n \), \( n^{-1} \log \Pr(\chi_4^2 > 2dn) \leq n^{-1} \log \Pr(\chi_4^2 > -2 \log E_{\Gamma_1, \Gamma_2|n} E_{\Gamma_2|n}) \leq n^{-1} \log \Pr(\chi_4^2 > 2cn) \). Therefore, \( -d \leq \liminf_{n \to \infty} n^{-1} \log E_{\Gamma_1, \Gamma_2|n} \leq \limsup_{n \to \infty} n^{-1} \log E_{\Gamma_1, \Gamma_2|n} \leq -c \). Let \( c, d \to (r_{1, \Gamma_1} + r_{2, \Gamma_2}) \) to get, \( \lim_{n \to \infty} n^{-1} \log E_{\Gamma_1, \Gamma_2|n} = -(r_{1, \Gamma_1} + r_{2, \Gamma_2}) \).

If, \( n^{-1} \log (1 - E_{2, \Gamma_2|n}) \to -a_{2, \Gamma_2} \) and (i) holds, then \( 2n^{-1} \log E_{\Gamma_1, \Gamma_2|n} E_{\Gamma_2|n} \to -2r_{1, \Gamma_1} \). The rest of the proof follows the same arguments as above.

\[ \blacksquare \]
2.5.3. Which evidence factor(s) provide evidence?

In an analysis based on evidence factors, it is useful if the decision to reject the null hypothesis can be attributed to one or both the factors. The closed testing principle of Marcus et al. (1976) can be used for this purpose. For a pair \((\Gamma_{1i}, \Gamma_{2i'})\), consider three comparisons:

(i) \(E_{\Gamma_{1i}, \Gamma_{2i'}} < \alpha\), (ii) \(E_{1, \Gamma_{1i}} < \alpha\), (iii) \(E_{2, \Gamma_{2i'}} < \alpha\). If (i), (ii), and (iii) are true, then we reject \(H_0\) based on evidence from both factors. If (i) and (ii) are true, then we reject based on the first factor. Similarly, if (i) and (iii) are true, then we reject based on the second factor. If only (i) is true, then rejection is based on the combined evidence alone, and the rejection decision cannot be attributed to one factor.

We are working in a scenario where it seems plausible that one assumption is true with the other being false. The following argument establishes that the above procedure preserves the probability of rejecting any \(\{H_{0, \Gamma_{1i}, \Gamma_{2i'}} \mid \Gamma_{1i} \geq \Gamma_1\} \cup \{H_{0, \Gamma_{1i}, \Gamma_{2i'}} \mid \Gamma_{2i} \geq \Gamma_2\}\) at the level \(\alpha\). If \(H_0\) is false there is nothing to prove. Assume \(H_0\) is true. Then, possible scenarios are: (1) \(H_0\) is true and both \(A_1(\Gamma_{1i})\) and \(A_2(\Gamma_{2i'})\) are true; (2) \(H_0\) is true and \(A_1(\Gamma_{1i})\) is true but \(A_2(\Gamma_{2i'})\) is false; and (3) \(H_0\) is true and \(A_2(\Gamma_{2i'})\) is true but \(A_1(\Gamma_{1i})\) is false. For any pair \((\Gamma_{1i}, \Gamma_{2i'})\) at most one of (1)–(3) can be true. When (1) holds, any false rejection implies \(\{E_{\Gamma_{1i}, \Gamma_{2i'}} < \alpha\}\), when (2) holds, a false rejection implies \(\{E_{1, \Gamma_{1i}} < \alpha\}\), and finally, when (3) holds, a false rejection implies \(\{E_{2, \Gamma_{2i'}} < \alpha\}\). Thus the familywise error rate is controlled at desired level \(\alpha\).

2.6. Simulation: combined evidence does better in finite sample

This section aims to verify that the Bahadur efficiency results of §2.5.2 provide an adequate guide to finite samples. We wish to verify that the combined evidence factor analysis requires a smaller sample size to make the correct decision with high probability than an analysis using either evidence factor alone.

Our simulation is based on the structure of the Life Span Study data (§2.2). We assume the data have \(S\) strata of triplets with exposures zero-dose, low-dose and high-dose. The
response is Bernoulli with probability \( \text{expit}(\alpha_s) \) if exposed zero-dose or \( \text{expit}(\alpha_s + \beta_l) \) if exposed to low-dose or \( \text{expit}(\alpha_s + \beta_h) \) for high-dose; here \( \text{expit}(x) = \frac{1}{1 + \exp(x)} \) and \( \beta_l \leq \beta_h \). The strata effect \( \alpha_s \) is sampled independently from \( N(0,0.2^2) \). The sample size \( (n) \) of §2.5.2 here is the number of strata \( (S) \), and increasing the sample size is equivalent to adding more and more strata while keeping the size of each stratum fixed. Figure 2 summarizes the simulation results in three panels of plots (A)–(C). Each panel corresponds to a separate simulation scenario with varied values of the effects \( \beta_l, \beta_h \). Within each panel three plots correspond to three different pairs of values of \( (\Gamma_1, \Gamma_2) \). Each plot shows the performance of the various tests as the sample size increases. Recall that \( \Gamma_1 \) is the sensitivity parameter for the high versus low dose comparison and \( \Gamma_2 \) is the sensitivity parameter for exposed versus unexposed comparison.

Panel (A) considers the null case \( \beta_l = \beta_h = 0 \). Recall that the simulation does not impose any bias in treatment assignment. In this situation, even a small amount bias will cause the probability of rejection to go to zero as the number of strata increases. A test is better if the rate at which this probability of rejection, plotted on the vertical axis, goes to zero is as fast as possible. For the graphs of panel (A), the higher the value, the faster, i.e. with less number of strata, we fail to reject the null on average. In plot A1, where \( \Gamma_1 = \Gamma_2 = 1.1 \), we see that as the number of strata increases the combined evidence narrowly beats both the factors. In plots A2 and A3 one of the two \( \Gamma_j \) values is large. The comparison with larger \( \Gamma_j \) always makes the correct decision; at least in the simulations. This is shown as a horizontal line at infinity. The plots show that the combined evidence dominates.

Panels (B) and (C) consider two scenarios under the alternative hypothesis: only high-dose has an effect, \( \beta_l = 0, \beta_h = 0.5 \), and both low-dose and high-dose has an effect, \( \beta_l = 0.5, \beta_h = 1 \), respectively. Here, in both the factors, as the number of strata increases the probability of acceptance will go to zero for bias below the design sensitivity. In these plots, the larger the graph is on the vertical axis plotting the rate of acceptance, the faster the null is rejected and the smaller the number of strata required to attain a certain power.
Figure 2: Graphs for combined evidence (solid), high to low dose (dashes) and exposed versus unexposed (dots). Panel (A) plots the negative of the rate of rejection of the hypothesis, \(-\log(\text{probability of rejection})/S\), in the null scenario, against the number of strata \(S\). Panels (B) and (C) plot the negative of the rate of acceptance of the hypothesis, \(-\log(\text{probability of acceptance})/S\) against \(S\). Along the rows \(\Gamma_1\) and \(\Gamma_2\) are varied. Results are based on average over 2000 simulations and over a grid of \(S\) values in gaps of 20.
In panel (B), the average (attributable) treatment effect in the comparison of exposed to unexposed units is considerably smaller. Consequently, the design sensitivity is smaller for the exposed versus unexposed comparison than that of the high versus low-dose comparison. Plot B1 considers bias levels $\Gamma_1 = 1.2$ and $\Gamma_2 = 1$. These $\Gamma$ values are chosen so that the power is not close to 0 or 1; otherwise we would not be able to compare methods clearly and get a sense of the rate based on the simulations. Correspondingly, as the effect in the comparison of high to low-dose is larger than that of exposed to unexposed comparison, $\Gamma_1$ is chosen to be larger than $\Gamma_2$. In this plot, the combined evidence dominates both the factors. For the next two plots, one of the two bias parameters are large so that the corresponding analysis is no longer able to detect the treatment effect with high power. Thus in these two scenarios the combined evidence borrows its strength mostly from only one factor. These plots show that as the number of strata increases, the rate for the combined evidence catches up with the better of the two factors.

Finally, in panel (C), the design sensitivity is smaller for the high versus low-dose comparison. The plots have similar behavior as in the plots of panel (B). Plot C1 of this panel considers the bias levels $\Gamma_1 = 1$ and $\Gamma_2 = 1.5$. The combined evidence has better performance compared to either of the factors. For the last two plots, as in panel (B), one of the bias parameters is taken to be large enough so that the corresponding analysis is no longer able to detect the treatment effect with high power. In both these scenarios we see the combined evidence has comparable performance to the better of the two factors.

2.7. Analysis of the Life Span Study data

The analysis to assess whether radiation has any carcinogenic effect consists of two comparisons, one based on comparing all proximal survivors with low and high doses and a second one comparing proximal survivors to not-in-city residents, giving us two evidence factors with $E_1 = 0.0021$ and $E_2 = 2.35 \times 10^{-10}$, see §2.2. The fact that these two comparisons form evidence factors is proved explicitly in the Appendix.
Figure 3: Result for testing radiation effect on solid cancer. In decreasing order of gradient the colors represent the decision - reject for both comparisons, for high to low dose comparison, for city to not-in-city comparison, without any attribution and do not reject. Bonferroni method rejects the null if $\Gamma_1 < 1.16$ and $\Gamma_2 < 1.11$ (dashed lines).
In the Life Span Study data, the observed confounders are age at exposure and sex. Then the bias levels $\Gamma_1$ and $\Gamma_2$ measure deviation from the assumptions that there is no unmeasured confounding for the comparisons of high-dose versus low-dose proximal survivors and all proximal survivors versus not-in-city residents, respectively, among individuals in the same strata of age at exposure and sex. The conclusion from the first comparison is sensitive at bias level $\Gamma_1 = 1.25$, i.e. we first fail to reject the hypothesis when $\Gamma_1 = 1.25$, whereas the conclusion from the second comparison is sensitive at bias level $\Gamma_2 = 1.12$. Therefore, to explain the observed associations, an unmeasured confounder, as in §2.2, would need to have a relatively weaker association with exposure to radiation in comparing all proximal survivors to not-in-city residents than in comparing high-dose to low-dose survivors. Figure 3 presents the results based on evidence factors. The factors are combined using the truncated product method with $\tilde{\alpha} = 0.20$ (see §2.4). The results show that the joint evidence is statistically significant for a carcinogenic effect for $(\Gamma_1, \Gamma_2) = (1.35, 1.12)$. However this decision cannot be attributed to either of the comparisons: at $(\Gamma_1, \Gamma_2) = (1.35, 1.12)$ each of the evidences considered separately are sensitive. At $(\Gamma_1, \Gamma_2) = (1.1, 1.3)$ the null hypothesis is rejected based on the evidence from comparison of proximal survivors with low and high doses. Another method to control for familywise error rate would be to use the Bonferroni correction. This leads to failing to reject the null for $\Gamma_1 \geq 1.16$ and $\Gamma_2 \geq 1.11$. Clearly, the Bonferroni method is conservative for small bias levels. For instance, at bias levels $(\Gamma_1, \Gamma_2) = (1.2, 1.13)$, we fail to reject the null after applying Bonferroni correction, but reject the null based on the joint evidence.

The sensitivity parameters, $\Gamma_1$ and $\Gamma_2$, models the biases in treatment assignment due to imbalance in unmeasured confounders. When calculating the evidence using this model, a near-perfect relationship is assumed between the unmeasured confounders and the response. It is not necessary to assume this near-perfect relationship – the one-parameter model with sensitivity parameter $\Gamma$ is equivalent to a set of models where with two sensitivity parameters: one that relates the unmeasured confounder to the response, $\Delta$; and one that relates the unmeasured confounder to the treatment, $\Lambda$. Rosenbaum and Silber (2009) show that
for each \( \Gamma \) in the one-parameter model, there is a curve of \( \Lambda \) and \( \Delta \) in this two-parameter model that gives equivalent inferences. For example, it follows that \( \Gamma = 1.25 \) is equivalent to an unobserved covariate that doubles the odds of treatment \( (\Lambda = 2) \) and doubles the odds of a positive treated-minus-control response difference \( (\Delta = 2) \). In the Appendix, we provide the technical discussion of this correspondence.

2.8. Discussion

Unmeasured confounding is a challenge in observational studies. Evidence factors, by constructing multiple independent sources of evidence that are potentially vulnerable to separate sources of unmeasured confounding, help us to either detect potential unmeasured confounding or make our findings more robust to unmeasured confounding. A practitioner might be concerned with loss of power from multiple comparisons when using evidence factors; this chapter establishes that if one constructs evidence factors and uses them carefully, as described in Theorem 3, there is no loss in power.

An alternative strategy in the Life Span Study could have been to select one of the two reference groups, distal survivors or non-in-city residents, and present a single analysis [French et al., 2017]. We showed that both reference groups can be used to build evidence factors. The combination of the factors provided evidence against the null hypothesis of no carcinogenic effect, which was robust to multiple sources of unmeasured confounding.

Our analysis was limited in the sense that it addressed whether there was a carcinogenic effect of radiation, but did not address the dose-response relationship. Currently, there is strong scientific interest in the shape of the dose-response curve, particularly at lower radiation doses, as well as differences in radiation risk by various demographic and lifestyle factors. The Life Span Study data contain rich individual level information that can be used to model these associations. Future research might seek to build evidence factors, perhaps by comparing survivors across multiple reasons of radiation exposure, to infer about the radiation dose-response and effect modification.
3.1. The need for approximation algorithms when constructing an observational design

3.1.1. Optimal constructions with polynomial-time algorithms

An algorithm is said to solve a problem in polynomial time if there is a polynomial, say \( n^3 \), such that the algorithm can solve any instance of the problem of size \( n \) in at most \( \kappa n^3 \) arithmetic operations, where \( \kappa \) is a constant, and in this case the algorithm runs in \( O(n^3) \)-time. Generally, the constant multiplier, \( \kappa \), depends upon the programming language, computer and other details, while the exponent, here 3, does not, so the focus of attention is on the exponent. Saying that a problem cannot be solved in polynomial time means saying that there are always problem instances such that it takes more than \( \kappa n^a \) arithmetic operations to solve them, no matter how the constants \( \kappa \) and \( a \) are picked. If a problem cannot be solved in polynomial time, then large problems may be virtually impossible to solve. See Papadimitriou and Steiglitz (1982) or Korte and Vygen (2012) for general discussion of combinatorial optimization algorithms and their performance. In this section, we first mention common uses of polynomial-time algorithms to construct observational studies, then point out that even some very simple and important designs cannot be built in this way.

Many methods of constructing a design for an observational study solve a combinatorial optimization problem using a polynomial-time algorithm. For instance, many treatment-control matching problems involving \( n \) people are reexpressed as optimal assignment problems or minimum cost flow problems in a network and are solved by algorithms that run in \( O(n^3) \)-time. In the simplest case, matched pairs are found to minimize the total over pairs of the covariate distance between the treated and control individuals within each pair; see, for instance, Rosenbaum (1989). The covariate distance might combine a propensity score with some form of Mahalanobis distance, or other techniques. The algorithm need not
construct pairs: it may construct matched sets with two controls matched to each treated individual, or it may be a full matching in which each matched set contains either one treated subject and one or more controls, or one control and one or more treated individuals; see, for instance, Hansen and Klopfer (2006). It is often useful to add fine-balance or near-fine balance constraints to minimum distance matching: these minimize the total distance within pairs subject to the constraint that a nominal variable, perhaps with many categories such as ICD-10 surgical procedure, is as balanced as possible; see, for instance, Rosenbaum, Ross and Silber (2007), Yang et al. (2012) and Pimentel, Yoon and Keele (2015) for methods and Silber et al. (2016) for an application. These techniques are implemented in R in Hansen’s optmatch package (Hansen and Klopfer 2006; Hansen, 2007) and Pimentel’s rcbalance package (Pimentel, 2016, 2017). Zubizarreta (2012) enlarges the scope of matching techniques in his designmatch package using mixed integer programming methods that often perform well despite lacking an explicit time bound (Zubizarreta and Kilcioglu 2016); Keele, Titiunik and Zubizarreta (2015) provide an application to enhancing regression discontinuity designs through matching.

A second polynomial-time algorithm used to design observational studies involves minimum distance nonbipartite matching; see Lu et al. (2011) and the references given there. Where bipartite (or two-part) matching pairs individuals from two groups, treated or control, nonbipartite (or, awkwardly, not-two-part) matching begins with a single population. Nonbipartite matching splits a single population into nonoverlapping matched pairs in such a way that the total distance within pairs is minimized. Derigs’ (1988) algorithm is available in the R package npMatching (Beck et al. 2016). One use of nonbipartite matching is in strengthening a weak instrument; see Baiocchi et al. (2010) and Keele and Morgan (2016) for discussion and Lorch et al. (2012) for an application.

For general discussion of matching in observational studies, see Rosenbaum (2010, Part II; 2017, Chapter 11) and Stuart (2010).
3.1.2. Closely related problems are much more difficult

If instead of matching two groups, treated and control, to minimize the distance within pairs, as in §3.1.1, we wished to match three groups of equal size to form minimum distance matched triples — the 3-dimensional assignment problem of Pierskalla (1968) — then the problem is believed to be very difficult and is classified among problems believed to have no solution by a polynomial-time algorithm; see Crama and Spieksma (1992, Theorem 1). Crama and Spieksma (1992, §3) proposed several approximation algorithms for matching everyone in three groups of initially equal size. These algorithms are not immediately applicable to statistical problems, because: (i) comparison groups are rarely of equal size prior to matching, (ii) we may want multiple controls from some groups, and (iii) one typically imposes additional constraints, such as fine-balance or near-fine balance for certain nominal variables; see §3.3.2. In §3.3, we employ ideas from Crama and Spieksma (1992) to produce a polynomial time approximation algorithm that incorporates features (i)-(iii).

Before discussing the algorithm, we motivate its use in an application in §3.2. For general discussion of approximation algorithms, see Vazirani (2010) and Williamson and Shmoys (2011).

3.2. Motivating Application: Effects of Side Airbags in Crashes

3.2.1. Side airbags: at first unavailable, then optional, then standard

According to the Insurance Institute for Highway Safety (http://www.iihs.org/iihs/ratings/safety-features), side airbags protecting both the driver’s head and torso were unavailable in 1997, and were standard equipment on 97.9% of cars in the 2017 model year. Over 20 years, more and more makes and models of cars gradually added side airbags, often offering them initially as optional equipment for an additional fee, later providing them as standard equipment. For instance, the Volvo C70 did not have side airbags in 1998, had them as optional equipment from 1999 to 2002, and then had them as standard equipment in 2003. The Nissan Altima did not have side airbags in 2006, had them as optional equipment from 2007-2009, and had them as standard equipment in 2010. It is not attractive to judge the
safety effects of side airbags by comparing a Volvo C70 to a Nissan Altima, because side
airbags are only one difference between these vehicles and their drivers. Perhaps Volvos
attract drivers concerned with safety, with the possible consequence that Volvos are driven
differently from, say, Dodge Chargers.

More attractive is to compare crashes of the same make and model of car in eras that differ
in availability of side airbags. For instance, one might compare a crash of a 1998 Volvo C70
to a crash of a 2003 Volvo C70, the latter having a side airbag. Presumably, the decision
to purchase a Volvo C70 in 1998 rather than in 2003 mostly reflects an individual’s need
to acquire or replace a car in those years, not a greater concern with automotive safety
by the purchaser of the 2003 model with side airbags. Nonetheless, the world changed in
many ways between 1998 and 2003, not just in the addition of side airbags to Volvo C70s,
so the wide separation in time raises other concerns. The comparison of the era before
side airbags and the era with side airbags as standard equipment is, in certain respects, an
attractive natural experiment, but it is not a perfect one.

In contrast, the middle era of optional side airbags has both attractions and concerns.
Models built in the same or adjacent years may be more similar than models separated by
five years. An attraction of the optional era is that one could compare a 2002 Volvo C70 to
a 2003 Volvo C70, or a 1998 Volvo C70 to a 1999 Volvo C70. Indeed, one could compare two
2000 Volvo C70’s, one with a side airbag, the other without. The concern is that, during
the optional period, 1999-2002 for Volvo C70s, some drivers paid extra for side airbags and
others declined to do so, perhaps indicating their different levels of concern with traffic
safety. Perhaps a driver concerned with automotive safety expresses that concern in more
than one way, say buying a side airbag, driving soberly and slowly, and abstaining from
tailgating, so that comparisons within the optional era confound side airbags with other
safety behavior. Presumably, the decision to purchase a Volvo C70 in 2000, rather than
1998 or 2003, again mostly reflects an individual’s need to acquire a car in 2000, but the
secondary decision, paying extra for side airbags in 2000, could be strongly related to other
unmeasured safe driving behaviors. So the optional era nudges people towards side airbags, but it does not determine whether they acquire a side airbag or not. One conventional strategy would compare the optional era to one of the other two eras, viewing the optional era as an instrument or instrumental variable for the purchase of a side airbag, thereby side-stepping the decisions of individual drivers about paying extra for a side airbag; see Baiocchi et al. (2014) for a review of instruments. Viewing the optional era in this way, an assumed exclusion restriction would attribute changes in injuries over eras to changes in the changed frequency of side airbags, leading to an estimate of the complier average effect of side airbags; see Angrist et al. (1996).

We will form matched sets consisting of crashes involving the same make and model of car, one before side airbags were available for this make and model, one after side airbags became standard equipment, and between 1 or 3 crashes in the optional period. As noted above for the Volvo C70 and Nissan Altima, the years involved vary from one make and model to another. We tried to find 3 crashes in the optional period because most buyers did not buy side airbags during the optional period. However, if there were too few crashes to form 1-3-1 matched sets, perhaps because the optional period was brief, we formed 1-1-1 matched triples instead. Ultimately, there were 978 matched sets of the form 1-3-1, and 1398 sets of the form 1-1-1. The data came from the US Fatality Analysis Reporting System (2017), described in §3.2.2.

3.2.2. The Fatality Analysis Reporting System
Provided by the US National Highway Traffic Safety Administration, the US Fatality Analysis Reporting System (FARS) records information about motor vehicle accidents with at least one fatality. The FARS data contain extensive information about the vehicles involved, some information about the occupants of the vehicles including a measure of severity of injury, 0 for uninjured to 4 for death, and some information about the circumstances of the crash. No doubt, crashes in FARS are unrepresentative of all crashes, because every crash in FARS was severe enough to cause at least one death.
Some care is needed when using FARS data to examine the effects of safety equipment. A crash is recorded only if there is at least one fatality. In FARS, a crash involving a lone driver hitting a tree is always lethal for the driver, not because driving alone is dangerous, nor because trees are invariably deadly, but because a crash involving just one person is recorded in FARS only if that person died. More subtly, if safety equipment prevents all deaths in a crash, then it also prevents the accident from being recorded in FARS, whereas removing the safety equipment might have caused a death, so the same crash would be recorded in FARS. For discussion of issues that arise when a treatment can cause data to go uncollected, see Rosenbaum (2005).

We looked at crashes involving at least two vehicles between 1995 and 2015. We included cars, minivans, SUVs and pickup trucks, but excluded motor cycles and large trucks. In such crashes, we picked at random one vehicle with at least one death, discarding that vehicle. The remaining vehicle or vehicles may or may not have had a death. We do know, however, that data on the remaining vehicles that we studied would have been collected by FARS whether or not side airbags or their absence caused or prevented a death in the vehicle, because the discarded vehicle would, in either case, have prompted FARS to collect data about this accident. This selection process makes the vehicles we examined unrepresentative of vehicles in FARS, but it avoids a tautological source of bias. We may hope that the selection process makes vehicles unrepresentative in a parallel manner in the three eras that we examine, the era prior to side airbags, the optional era, and the era when they were standard equipment. For brevity in tables and figures, these three eras are called “none”, “optional” and “all”, and by definition the years involved vary from one make and model to another. We considered only makes and models that had cars in all three eras, so that, for instance, we would exclude a new make of car that was first sold in 2012 with standard side airbags.

Many car models had no cars in one or more eras, “none”, “optional” and “all”. For instance, a new car model might have had side airbags from the beginning. A discontinued
car model might never have had side airbags. A car model might go from “none” to “all” with no “optional” period. We required a car model to have at least 40 cars in each of the three periods, “none”, “optional” and “all”, and we used these 40 cars to define the eras. There were 31,505 cars that qualified. This number was slightly reduced by 2,282 cars due to missing data on key variables. For each car model, we matched the smallest group, “none”, “optional” or “all”, to the two larger groups. Where possible, we matched 1-3-1, none-optional-all, because most cars in the optional period did not have side airbags, so selecting three cars increased the chance that one had side airbags. If 1-3-1 was not possible because 3 optional era cars were not available, we matched 1-1-1. This yielded 2376 matched sets. In the end, we had 978 matched 1-3-1 sets and 1398 matched 1-1-1 sets, where 2376 = 1398 + 978, with a total of 9084 cars in these 2376 matched sets.

The matching used the new algorithm in §3.3. The R package approxmatch implements the procedure in §3.3; Karmaker (2017; R Core Team 2018). That package includes a data frame called Dodgeram with 6953 crashes involving Dodge Ram trucks. The examples in the documentation for the kwaymatching function in the approxmatch package in R reproduce the 3-way matching of the Dodge Ram trucks.

3.2.3. Matched crashes

Table 3 describes covariate balance after matching. In Table 3, each matched set counts the same, so the optional period for a 1-1-1 triple contributes one driver age for the optional period, but a 1-3-1 matched set contributes one average of three driver ages. Table 3 shows characteristics of the driver, such as age, and of the crash, such the direction of impact and whether a fire or explosion occurred. By definition, we did not match for the model year, nor did we match for the crash year. The years are out of balance by definition: for each make and model, the none-era precedes the optional era which precedes the all-era. In the optional era, about 18% of owners had purchased cars with side airbags. We also matched for some additional variables not shown in Table 3, such as characteristics of the right front passenger if there was one, and the stated highway speed limit which may or may not have
been heeded.

Table 3 shows the balance achieved by matching, but it does not contrast the situations before and after matching. Figures 4 and 5 show this contrast, with the left bar showing the situation before matching, the right bar showing the situation after matching. We hope to see that the right bars, after matching, are of similar height, and indeed they are. Plots use \texttt{ggplot2} (Wickham 2009). Notably in Figure 4, belt use and age are similar after matching, but they were different before matching. The mean driver’s age was about four years younger in the “none” period than in the “all” period, perhaps because the baby-boomers are aging, and there was also about a 12% increase in use of safety belts over this period. Also, there were more female drivers in the later “all” period. Of course, the match controlled for age, gender and restraint use, but there could be other factors that were not measured. Notably in Figure 5, the direction of impact shifted slightly as the periods passed, with an increase in rear and right impacts, and a decrease in front impacts and rollovers. Recall that FARS records only lethal crashes, so this is a change in the pattern of lethal crashes, not necessarily a change in the pattern of all impacts. Again, the matching removed these differences.

Table 3: Balance on covariates in 1398 matched 1-1-1 matched triples and 978 matched 1-3-1 sets. Cars were also matched for make and model. Group “none” refers to an era when side airbags were not available for this make/model, “optional” to an era when side airbags were optional for this make/model, and “all” to an era when all cars of this make/model had side airbags. For driver’s age, crash year, and model year, values are means; otherwise, they are percentages. A mean is computed within a matched set, then averaged over sets.

<table>
<thead>
<tr>
<th>Group</th>
<th>Driver</th>
<th>Direction of Impact</th>
<th>Roll-over</th>
<th>Fire</th>
<th>Year</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Age</td>
<td>Female</td>
<td>Belted</td>
<td>Left</td>
<td>Right</td>
</tr>
<tr>
<td>None</td>
<td>40</td>
<td>35</td>
<td>87</td>
<td>14</td>
<td>13</td>
</tr>
<tr>
<td>Optional</td>
<td>40</td>
<td>37</td>
<td>88</td>
<td>14</td>
<td>13</td>
</tr>
<tr>
<td>All</td>
<td>41</td>
<td>37</td>
<td>87</td>
<td>14</td>
<td>12</td>
</tr>
</tbody>
</table>

The match used a distance that satisfied the triangle inequality, and it finely balanced several covariates and was exact for the presence of a right front passenger. Specifically, the match near-finely balanced indicators of rollover and fire occurrence during the crash. Following the idea in §3.4, the distance was the weighted sum of distances involving the
Figure 4: Driver’s characteristics before matching and balance of the characteristics after matching across the three eras, None, Optional, and All. Bars of similar height after matching indicate that matching has balanced the covariate.
Figure 5: Characteristics of the crash before matching and balance of the characteristics after matching across the three eras, None, Optional, and All. Bars of similar height after matching indicate that matching has balanced the covariate.
absolute difference in logits of the propensity score, a rank based Mahalanobis distances for
occupant characteristics, and two other rank based Mahalanobis distances for direction of
impact and safety belt use; see the examples in the `approxmatch` package in R. Weighting
several distances permits the combined distance to satisfy the triangle inequality while giving
the analyst control of the relative importance of variables in the match. In principle,
finding an optimal 1-1-1 match or 1-3-1 match is for all intents impossible in large problems.
The match was produced using the approximation algorithm we describe and develop in
§3.3. The approximation algorithm runs in polynomial time.

3.3. Problem, algorithm, and guarantee

3.3.1. Matching structure and distance
There are three disjoint sets of units, \( I = \{1, \ldots, I\} \), \( J = \{1, \ldots, J\} \) and \( K = \{1, \ldots, K\} \)
with \( I \leq J \leq K \). In §3.2, the sets \( I \), \( J \), and \( K \) were eligible car crashes in the three eras,
“none”, “optional” and “all”.

There is a distance, \( \delta_{ij} \geq 0 \), between pairs of units, \( i \in I \) and \( j \in J \), a distance \( \delta'_{ik} \geq 0 \)
between pairs of units in \( i \in I \) and \( k \in K \), and a distance \( \delta''_{jk} \geq 0 \) between pairs of units
in \( j \in J \) and \( k \in K \). Write \( \delta \) for the vector of \( IJ + IK + JK \) distances, \( \delta_{ij}, \delta'_{ik}, \delta''_{jk} \),
i = 1, \ldots, I, j = 1, \ldots, J, k = 1, \ldots, K. Although it is suggestive to call the numbers in \( \delta \)
distances, they are required to have some but not all of the properties of distances in a metric
space. Precisely, we require without further mention that the entries in \( \delta \) be nonnegative,
possibly infinite, and satisfy a part of the triangle inequality, namely \( \delta''_{jk} \leq \delta_{ij} + \delta'_{ik} \), and
we call such a \( \delta \) a “matching distance array”, or briefly a “distance”. Notice that the
triangle inequality bounds distances \( \delta''_{jk} \) between units in \( j \in J \) and \( k \in K \), but need not
bound \( \delta_{ij} \) for \( i \in I \) and \( j \in J \), nor \( \delta'_{ik} \) for \( i \in I \) and \( k \in K \). We need \( \delta''_{jk} \leq \delta_{ij} + \delta'_{ik} \), but
not other implications of the triangle inequality, because our algorithm makes \( \delta_{ij} \) and \( \delta'_{ik} \)
small, then concludes that \( \delta''_{jk} \) could not be very big by virtue of this inequality. Distances
are not defined for two units in \( I \), nor for two units in \( J \), nor for two units in \( K \); rather,
distances are defined between units in different sets. We understand \( \delta''_{jk} \leq \delta_{ij} + \delta'_{ik} \) to
hold trivially if either $\delta_{ij} = \infty$ or $\delta'_{ik} = \infty$. For instance, if individual $i$ has covariates $x_i$, individual $j$ has covariates $x_j$, and $\Sigma$ is the covariance matrix of the covariates, then the Mahalanobis distance defined as $\delta_{ij} = \sqrt{(x_i - x_j)^T \Sigma^{-1} (x_i - x_j)}$ satisfies the triangle inequality. Although the Mahalanobis distance yields a formula, we do not assume that a single formula produced the $\delta_{ij}$, $\delta'_{ik}$, $\delta''_{jk}$, leaving open the possibility that $\delta''_{jk}$ is defined differently from $\delta_{ij}$ or $\delta'_{ik}$, requiring only that $\delta''_{jk} \leq \delta_{ij} + \delta'_{ik}$. Section 3.4 discusses various useful distances that satisfy the required conditions.

Let $\rho \geq 1$ be an integer such that $\rho \leq J/I$ and let $\kappa \geq 1$ be an integer such that $\kappa \leq K/I$. A common and basic case is $(\rho, \kappa) = (1, 1)$. We want to construct a closely matched and balanced blocked study such that each block contains 1 unit from $I$, $\rho$ units from $J$ and $\kappa$ units from $K$, and no units appear more than once. If $(\rho, \kappa) = (1, 1)$, then we want matched triples with one unit each from $I$, $J$ and $K$.

More precisely, a $(1 + \rho + \kappa)$-tuple $B = (i, j_1, j_2, \ldots, j_\rho, k_1, k_2, \ldots, k_\kappa)$ is a $(\rho, \kappa)$-block if $i \in I$, $j_1 \in J$, $j_\rho \in J$, $k_1 \in K$, $k_\kappa \in K$ where $j_1, \ldots, j_\rho$ are distinct and $k_1, \ldots, k_\kappa$ are distinct. A $(\rho, \kappa)$-design $B$ is a collection of blocks $B$ such that every unit $i \in I$ appears in exactly one block, and each $j \in J$ and each $k \in K$ appears in at most one block $B \in B$.

A block $B = (i, j_1, j_2, \ldots, j_\rho, k_1, k_2, \ldots, k_\kappa)$ has total between group distance

$$\delta_B = \sum_{\ell=1}^{\rho} \delta_{i,j_\ell} + \sum_{m=1}^{\kappa} \delta'_{i,k_m} + \sum_{\ell=1}^{\rho} \sum_{m=1}^{\kappa} \delta''_{j_\ell,k_m}. \quad (3.1)$$

If $\delta_B$ is very small, then $i$ is typically close to $j_1, j_2, \ldots, j_\rho$ and close to $k_1, k_2, \ldots, k_\kappa$, each $j$ in the block is close to each $k$ in the block. The design $B$ has total distance $\delta_B = \sum_{B \in B} \delta_B$.

We would prefer a design in which $\delta_B$ is small.

The distance in (3.1) worries about distances between $I$, $J$, and $K$, not distances within them. To see why this is reasonable, consider a simple case. Suppose that $\delta_{ij}$ is the absolute
difference in age and $\rho = 2$. If $i$ has age 25, $j_1$ has age 24 and $j_2$ has age 26, then $\sum_{\ell=1}^{\rho} \delta_{i,j_{\ell}}$ in (3.1) is $|25 - 24| + |25 - 26| = 2$. In contrast, if $i$ is 24, $j_1$ is 25 and $j_2$ is 26, then $\sum_{\ell=1}^{\rho} \delta_{i,j_{\ell}}$ is $|24 - 25| + |24 - 26| = 3$. So we prefer the first triple of ages to the second, and this makes sense because in the second distribution of ages $i$ is younger than both $j_1$ and $j_2$, so they are inferior as a control group. Had the first term in (3.1) included a distance between the two units from $J$, with three terms instead of $\delta_{i,j_1} + \delta_{i,j_2}$ in the first sum in (3.1), then $|25 - 24| + |25 - 26| + |24 - 26| = 4$ for both distributions of age, so the distance would not represent our preference for the first distribution of ages.

3.3.2. Fine balance and near-fine balance of nominal categories

Fine balance entails equating, in two or more groups, the marginal distributions of a nominal covariate, often a covariate with many levels, without worrying about whether individuals are paired for this covariate. A very large completely randomized experiment balances the marginal distributions of covariates without pairing individuals, and fine balance aims at an analogous form of balance for an observed nominal covariate. Fine balance can ensure that many nominal categories are balanced, while permitting the pairing to focus on other covariates strongly associated with the outcome.

There are $C \geq 1$ mutually exclusive and exhaustive categories, $C_1, \ldots, C_C$, such that every unit, $i \in I$ belongs to exactly one category, say $i \in C_c$ for one specific $c$, and the same is true for every unit, $j \in J$ and every unit $k \in K$. For example, $C_1$ might be the set of females and $C_2$ might be the set of males. More commonly, $C_1, \ldots, C_C$ might represent dozens or hundreds of categories, say principal surgical procedures or car models. The trivial case $C = 1$ places everyone in the same category, and it will permit a single theorem to cover the situations with categories ($C > 1$) and without categories ($C = 1$). Write $f_{cg}$ for the number of units in category $c$ from group $g = 1, 2, 3$, where group 1 is $I$, group 2 is $J$ and group 3 is $K$.

If $B$ is a $(\rho, \kappa)$-design, then it exhibits a certain degree of imbalance with respect to the
categories $C_1, \ldots, C_C$. Write $f_{B}$ for a $C \times 3$ matrix of counts defined as follows. For $(\rho, \kappa)$-design $B$, the count, $f_{Bcg}$ in row $c$ and column $g$ is the number of units in $B$ in category $c$ from group $g$. So, by definition, the column totals are $\sum_{c=1}^{C} f_{Bc1} = I$, $\sum_{c=1}^{C} f_{Bc2} = \rho I$ and $\sum_{c=1}^{C} f_{Bc3} = \kappa I$. Because $B$ takes everyone in group $I$, we have $f_{c1} = f_{Bc1}$ for each $c$, and because $B$ takes some people from groups $J$ and $K$, we have $f_{c2} \geq f_{Bc2}$ and $f_{c3} \geq f_{Bc3}$.

The $(\rho, \kappa)$-design $B$ is finely balanced if $f_{Bc1} = f_{Bc2}/\rho = f_{Bc3}/\kappa$ for $c = 1, \ldots, C$; that is, every matched group has the same proportion of individuals in each category. Fine balance is discussed in Rosenbaum (1989, §3.2; Rosenbaum 2010, §10). Fine balance is not always feasible. Write $f_{c\text{min}} = \min(f_{c1}, f_{c2}/\rho, f_{c3}/\kappa)$. If $f_{c\text{min}} = f_{c1}$ for $c = 1, \ldots, C$, then fine balance is feasible, and if it is feasible, then we wish to require it. The $(\rho, \kappa)$-design $B$ is near-fine or exhibits near-fine balance if $f_{Bc1} \geq f_{c\text{min}}$, $f_{Bc2}/\rho \geq f_{c\text{min}}$ and $f_{Bc3}/\kappa \geq f_{c\text{min}}$ for $c = 1, \ldots, C$. Near-fine balance is always feasible, and when fine balance is feasible, near-fine balance implies fine-balance. In a sense, near-fine balance exhibits minimal deviation from fine balance. Requiring near-fine balance when $C = 1$ imposes no constraint, and in this case the focus is entirely on minimizing distance, $\delta_B$, with no concern for balance over categories. This definition of near-fine balance is similar to the definition in Yang et al. (2012) but has been adjusted to permit three groups instead of a treated and a control group.

The traditional three-dimensional assignment problem is to minimize $\delta_B$ with $C = 1$, $\rho = \kappa = 1$, and $I = J = K$, and even in this simplest case, finding an optimal solution is not practical; see Crama and Spieksma (1992). Consider the general problem of finding a near-fine $(\rho, \kappa)$-design $B$ with a small distance, $\delta_B$. We propose an approximation algorithm for general $C$, $\rho$, $\kappa$, $I$, $J$, $K$: it finds a near-fine $(\rho, \kappa)$-design with a total distance, $\delta_B$, that is at most $1 + \max(\rho, \kappa)$ times the minimum distance for all near-fine $(\rho, \kappa)$-designs. The approximation algorithm runs in $O(K^3)$ time in the worst case, and for matched triples, $\rho = \kappa = 1$, it produces a value of $\delta_B$ that is at most twice the true minimum as $1 + \max(\rho, \kappa) = 2$. 

39
3.3.3. An approximation algorithm

Define a partial block \( P = (i, j_1, j_2, \ldots, j_\rho) \) to be \( \rho + 1 \) distinct units with \( i \in I \) and \( j_1, j_2, \ldots, j_\rho \in J \). A set \( \mathcal{P} \) of partial blocks will be called acceptable if it is the initial part of some near-fine \((\rho, \kappa)\)-design \( \mathcal{B} \); however, this needs to be said with a bit more care.

A set \( \mathcal{P} \) of partial blocks \( P \) is compatible with a near-fine \((\rho, \kappa)\)-design \( \mathcal{B} \) if each block \( B = (i, j_1, j_2, \ldots, j_\rho, k_1, k_2, \ldots, k_\kappa) \in \mathcal{B} \) has an initial segment \( P = (i, j_1, j_2, \ldots, j_\rho) \) that is a partial block in \( \mathcal{P} \). A set \( \mathcal{P} \) of partial blocks is defined to be acceptable if its partial blocks are compatible with at least one near fine \((\rho, \kappa)\)-design \( \mathcal{B} \). If \( \mathcal{P} \) is acceptable, then it necessarily follows that its partial blocks are nonoverlapping and \( f_{B_1} \geq f_{c_{\min}} \), \( f_{B_2}/\rho \geq f_{c_{\min}} \). An unexciting but nonetheless important subtlety here is that \( f_{B_1} \) and \( f_{B_2} \) are determined by \( \mathcal{P} \) without reference to group \( K \), but \( f_{c_{\min}} \) is defined in a way that involves group \( K \).

The following two-step algorithm first assigns \( j \)’s in \( J \) to each \( i \in I \) to produce an acceptable set \( \mathcal{P} \) of partial blocks \( P = (i, j_1, j_2, \ldots, j_\rho) \); then, it assigns \( k \)’s in \( K \) to each partial block, \( P \).

**Step 1:** Match \( I \) and \( J \) to form an acceptable set \( \mathcal{P} \) of partial blocks of minimal distance

\[
\sum_{(i, j_1, j_2, \ldots, j_\rho) \in \mathcal{P}} \sum_{\ell=1}^{\rho} \delta_{i,j_\ell}.
\] (3.2)

**Step 2:** Extend each partial block \( P = (i, j_1, j_2, \ldots, j_\rho) \in \mathcal{P} \) into a block

\[
B = (i, j_1, j_2, \ldots, j_\rho, k_1, k_2, \ldots, k_\kappa)
\]

so that the resulting collection of blocks \( \mathcal{B} \) is a near-fine \((\rho, \kappa)\)-design \( \mathcal{B} \) that minimizes

\[
\sum_{(i, j_1, j_2, \ldots, j_\rho) \in \mathcal{P}} \sum_{m=1}^{\kappa} \left( \delta'_{i,k_m} + \sum_{\ell=1}^{\rho} \delta''_{j_\ell,k_m} \right).
\]
Step 1 matches individuals in $I$ to individuals in $J$ forming partial blocks, while Step 2 takes those partial blocks and matches each partial block to individuals in $K$. In other words, the entire procedure consists of two matches, one of individuals to individuals, the other of partial blocks to individuals. The entire procedure is suboptimal because Step 1 makes decisions with no allowance for their consequences in Step 2, but we will show that the procedure’s mistakes are limited in size. It is the triangle inequality that limits the size of the errors.

Consider the requirement of near-fine balance, namely the requirement that $f_{Bc1} \geq f_{c\text{min}}$, $f_{Bc2}/\rho \geq f_{c\text{min}}$, and $f_{Bc1}/\kappa \geq f_{c\text{min}}$. This definition of near-fine balance refers to all three groups because $f_{c\text{min}}$ is derived from all three groups. The definition of $f_{c\text{min}}$ ensures that the requirement of near-fine balance is feasible: a $(\rho, \kappa)$-design $B$ exhibiting near-fine balance always exists, albeit perhaps with an infinite total distance if some distances $\delta$ are infinite. Because Step 1 requires the partial blocks to be acceptable, they are compatible with near-fine balance; that is, the partial blocks are the initial parts of the blocks of some $(\rho, \kappa)$ design $B$ that exhibits near-fine balance. Step 2 requires that these partial blocks be extended to complete blocks so that the resulting $(\rho, \kappa)$-design $B$ exhibits near-fine balance. So, by the definitions of Steps 1 and 2, the algorithm returns a $(\rho, \kappa)$-design $B$ exhibiting near-fine balance, albeit perhaps with an infinite total distance.

Can Steps 1 and 2 be performed? Indeed they can, and in computational time that is $O(K^3)$. Step 1 can be done in $O(K^3)$ steps by solving a minimum cost network flow problem, matching elements of $I$ to $\rho$ elements of $J$ with a requirement of near-fine balance defined by the given values $f_{c\text{min}}$; see Rosenbaum (1989, §3.2) or Yang et al. (2012, Appendix) with very minor changes to accommodate the value $f_{c\text{min}}$ obtained from all three groups. (Specifically, in the language of these references, the capacity of the edge from the category $c$ node to the sink is set to $\rho f_{c\text{min}}$ in Step 1 and set to $\kappa f_{c\text{min}}$ in Step 2, with one additional bypass node with capacity $\rho I - \rho \sum f_{c\text{min}}$ in Step 1 or capacity $\kappa I - \kappa \sum f_{c\text{min}}$ in Step 2 delivering the remaining flow to the sink.) The match in Step 2 can also be done in
$O(K^3)$ steps by solving a minimum cost network flow problem, now matching each $P \in \mathcal{P}$ to $\kappa$ elements of $\mathcal{K}$, so that the selected controls from $\mathcal{K}$ again exhibit near-fine balance defined by the given values of $f_{c_{\min}}$.

In brief, the approximation algorithm entails solving two minimum cost flow problems, each of which runs in $O(K^3)$ steps, so the approximation algorithm itself is a polynomial time algorithm that runs in $O(K^3)$ steps.

3.3.4. A bound on the error of approximation

Proposition 1 says that the polynomial-time algorithm in §3.3.3 finds a near-fine design $\mathcal{B}$ whose total distance $\delta_{\mathcal{B}}$ is at most a fixed multiple of the distance $\delta_{\mathcal{B}^*}$ for the unattainable optimal near-fine design, say $\mathcal{B}^*$. The multiplier, $1 + \max(\rho, \kappa)$, in Proposition 1 equals 2 in the common case with $\rho = \kappa = 1$.

The proof of Proposition 1 extends certain ideas from Crama and Spieksma (1992) for the simpler 3-dimensional assignment problem to matching with (i) near-fine balance, (ii) unequal initial groups, $I \leq J \leq K$, and (iii) matching with multiple controls, $\rho \geq 1$ and $\kappa \geq 1$. One device they use with $I = J = K$ cannot be used here: Step 1 of our algorithm must start with the smallest group, $\mathcal{I}$.

**Proposition 1** Let $\overline{\mathcal{B}}$ be a near fine $(\rho, \kappa)$-design whose distance $\delta_{\overline{\mathcal{B}}}$ is minimal. Let $\mathcal{B}$ be produced by Steps 1 and 2 in §3.3.3. Then $\mathcal{B}$ is a near-fine $(\rho, \kappa)$-design with $\delta_{\mathcal{B}} \leq \{1 + \max(\rho, \kappa)\} \delta_{\overline{\mathcal{B}}}$.

**Proof.** Let $\mathcal{P}$ be the set of partial blocks $P = (i, j_1, j_2, \ldots, j_\rho)$ in $\mathcal{B}$ produced in Step 1. We now define a compromise $\tilde{\mathcal{B}}$ between $\mathcal{B}$ and $\overline{\mathcal{B}}$, anchoring the compromise by units $i \in \mathcal{I}$. Define $\tilde{\mathcal{B}}$ to be a $(\rho, \kappa)$-design with near-fine balance whose blocks $\tilde{B} = (i, j_1, j_2, \ldots, j_\rho, k_1, \ldots, k_\kappa)$ are such that $P = (i, j_1, j_2, \ldots, j_\rho) \in \mathcal{P}$ is a partial block of $\mathcal{P}$, and for some $\ell_1, \ldots, \ell_\rho \in \mathcal{J}$, $\overline{B} = (i, \ell_1, \ldots, \ell_\rho, k_1, \ldots, k_\kappa) \in \overline{\mathcal{B}}$ is a block of the
optimal design. By the triangle inequality, we always have

$$
\delta_{j_\ell, k_m}'' \leq \delta_{i,j_\ell} + \delta_{i,k_m}'
$$

(3.3)

By definition,

$$
\delta_B = \sum_{(i, j_1, j_2, \ldots, j_\rho, k_1, \ldots, k_\kappa) \in B} \left( \sum_{\ell=1}^{\rho} \delta_{i,j_\ell} + \sum_{m=1}^{\kappa} \delta_{i,k_m}' + \sum_{\ell=1}^{\rho} \sum_{m=1}^{\kappa} \delta_{j_\ell,k_m}'' \right).
$$

Now $B$ and $\bar{B}$ have the same partial blocks, $P = (i, j_1, j_2, \ldots, j_\rho)$, but for these fixed partial blocks, Step 2 completed the partial blocks in $B$ as $(i, j_1, j_2, \ldots, j_\rho, k_1, \ldots, k_\kappa)$ where the $k$’s were chosen to minimize $\sum (i, j_1, j_2, \ldots, j_\rho, k_1, \ldots, k_\kappa) \in B \left( \sum_{m=1}^{\kappa} \delta_{i,k_m}' + \sum_{\ell=1}^{\rho} \sum_{m=1}^{\kappa} \delta_{j_\ell,k_m}'' \right)$; therefore,

$$
\delta_B \leq \delta_{\bar{B}} = \sum_{(i, j_1, j_2, \ldots, j_\rho, k_1, \ldots, k_\kappa) \in \bar{B}} \left( \sum_{\ell=1}^{\rho} \delta_{i,j_\ell} + \sum_{m=1}^{\kappa} \delta_{i,k_m}' + \sum_{\ell=1}^{\rho} \sum_{m=1}^{\kappa} \delta_{j_\ell,k_m}'' \right).
$$

(3.4)

Applying the triangle inequality (3.3) to $\sum_{\ell=1}^{\rho} \sum_{m=1}^{\kappa} \delta_{j_\ell,k_m}''$ in (3.4) yields

$$
\delta_B \leq \delta_{\bar{B}} \leq \sum_{(i, j_1, j_2, \ldots, j_\rho, k_1, \ldots, k_\kappa) \in \bar{B}} \left\{ (1 + \kappa) \sum_{\ell=1}^{\rho} \delta_{i,j_\ell} + (1 + \rho) \sum_{m=1}^{\kappa} \delta_{i,k_m}' \right\}
$$

$$
\leq \{1 + \max(\rho, \kappa)\} \sum_{(i, j_1, j_2, \ldots, j_\rho, k_1, \ldots, k_\kappa) \in \bar{B}} \left\{ \sum_{\ell=1}^{\rho} \delta_{i,j_\ell} + \sum_{m=1}^{\kappa} \delta_{i,k_m}' \right\}.
$$

(3.5)

Because $B$ and $\bar{B}$ have the same partial blocks, $P = (i, j_1, j_2, \ldots, j_\rho)$, and Step 1 picked the $(j_1, j_2, \ldots, j_\rho)$ in these blocks to minimize $\sum (i, j_1, j_2, \ldots, j_\rho, k_1, \ldots, k_\kappa) \in B \sum_{\ell=1}^{\rho} \delta_{i,j_\ell},$

$$
\sum_{(i, j_1, j_2, \ldots, j_\rho, k_1, \ldots, k_\kappa) \in \bar{B}} \sum_{\ell=1}^{\rho} \delta_{i,j_\ell} = \sum_{(i, j_1, j_2, \ldots, j_\rho, k_1, \ldots, k_\kappa) \in B} \sum_{\ell=1}^{\rho} \delta_{i,j_\ell}
\leq \sum_{(i, j_1, j_2, \ldots, j_\rho, k_1, \ldots, k_\kappa) \in \bar{B}} \sum_{\ell=1}^{\rho} \delta_{i,j_\ell}.
$$

(3.6)
By the definition of $\tilde{B}$,

$$
\sum_{(i, j_1, j_2, ..., j_\rho, k_1, ..., k_\kappa) \in \tilde{B}} \sum_{m=1}^{\kappa} \delta'_{i,k_m} = \sum_{(i, j_1, j_2, ..., j_\rho, k_1, ..., k_\kappa) \in \tilde{B}} \sum_{m=1}^{\kappa} \delta'_{i,k_m}.
$$

(3.7)

Combining (3.5), (3.6) and (3.7) yields

$$
\delta_B \leq \{1 + \max (\rho, \kappa)\} \sum_{(i, j_1, j_2, ..., j_\rho, k_1, ..., k_\kappa) \in \tilde{B}} \left\{ \sum_{\ell=1}^{\rho} \delta_{i,j\ell} + \sum_{m=1}^{\kappa} \delta'_{i,k_m} \right\}
$$

$$
\leq \{1 + \max (\rho, \kappa)\} \sum_{(i, j_1, j_2, ..., j_\rho, k_1, ..., k_\kappa) \in \tilde{B}} \left\{ \sum_{\ell=1}^{\rho} \delta_{i,j\ell} + \sum_{m=1}^{\kappa} \delta'_{i,k_m} + \sum_{\ell=1}^{\rho} \sum_{m=1}^{\kappa} \delta''_{j\ell,k_m} \right\}
$$

$$
= \{1 + \max (\rho, \kappa)\} \delta_{\tilde{B}}.
$$

3.3.5. The bound is tight

For $\rho = \kappa = 1$, or 1-1-1 matching, the bound in Proposition 1 is $\delta_B \leq \{1 + \max (\rho, \kappa)\} \delta_{\tilde{B}} = 2 \delta_{\tilde{B}}$. This bound cannot be improved without additional assumptions. To see this, consider $\mathcal{I} = \{i\}$, $\mathcal{J} = \{j_1, j_2\}$, $\mathcal{K} = \{k_1, k_2\}$, where $x_i = 1$, $x_{j_1} = x_{k_1} = 0$, $x_{j_2} = 2 - \epsilon$, and $x_{k_2} = 3 - 2\epsilon$, where $0.01 > \epsilon > 0$, and the distance between $a$ and $b$ is $|x_a - x_b|$, as depicted in (3.8).

$$
\begin{array}{|c|c|c|c|}
\hline
x & 0 & -1 & (2 - \epsilon) & (3 - 2\epsilon) \\
unit & j_1, k_1 & i & j_2 & k_2 \\
\hline
\end{array}
$$

(3.8)

The optimal 1-1-1 match $\overline{B}$ is $(i, j_1, k_1)$ with $\delta_{\overline{B}} = |x_i - x_{j_1}| + |x_i - x_{k_1}| + |x_{j_1} - x_{k_1}| = 1 + 1 + 0 = 2$. The approximation algorithm would first pair $(i, j_2)$ with a distance of $|x_{j_2} - x_i| = |2 - \epsilon - 1| = 1 - \epsilon < 1 = |x_{j_1} - x_i|$. Then, the approximation algorithm would pair $(i, j_2)$ with $k_2$ rather than with $k_1$ because $|x_i - x_{k_2}| + |x_{j_2} - x_{k_2}| = (2 - 2\epsilon) + (1 - \epsilon) = 3 - 3\epsilon < |x_i - x_{k_1}| + |x_{j_2} - x_{k_1}| = 1 + 2 - \epsilon = 3 - \epsilon$. So the approximation algorithm would yield $B$ consisting $(i, j_2, k_2)$ with distance $\delta_B = |x_{j_2} - x_i| + |x_{k_2} - x_i| + |x_{j_2} - x_{k_2}| =$
\[ |1 - \epsilon| + |2 - 2\epsilon| + |1 - \epsilon| = 4 - 4\epsilon. \] Because \( \epsilon > 0 \) can be made arbitrarily small, the best bound is \( B = 2B \). Problems with \( B = 2B \) of any size can be constructed by replicating this example with \( x \)'s that are spaced apart by, say, 10 units for each replicate, say at 10, 11, 12 - \( \epsilon \) and 13 - 2\( \epsilon \) for the second replicate.

3.4. Covariate distances that satisfy the needed triangle inequality

Statistical matching has often used covariate distances, but is typically indifferent about whether those distances satisfy the triangle inequality. Recall that a matching distance array \( \delta \) only requires nonnegative, possibly infinite numbers such that \( \delta_{jk} \leq \delta_{ij} + \delta_{ik} \), \( i = 1, \ldots, I, \ j = 1, \ldots, J, \ k = 1, \ldots, K \), with some convenient consequences when \( \delta_{jk} \) is defined differently from \( \delta_{ij} \) and \( \delta_{ik} \). In this section, we briefly discuss some options in defining \( \delta \).

If \( \delta \) and \( \tilde{\delta} \) are two matching distance arrays, then so is \( w\delta + \tilde{w}\tilde{\delta} \) for any nonnegative real numbers \( w \geq 0, \tilde{w} \geq 0 \). For instance, a positively weighted combination of two or more Mahalanobis distances involving different, perhaps overlapping, sets of covariates, yields a new distance array. This permits some covariates to receive greater emphasis, others to receive less. If a first match using distances \( \delta \) leaves an unsatisfactory imbalance for a few covariates, then use of a second distance \( w\delta + \tilde{w}\tilde{\delta} \) may fix the problem if \( \tilde{\delta} \) emphasizes the problematic covariates.

With three groups, we may estimate a two-dimensional propensity score; see Imai and van Dyk (2004). A matching distance array is obtained as a positively weighted combination of a Mahalanobis distance for the two dimensional propensity score and a Mahalanobis distance for all or for a subset of covariates used in defining the scores.

A robust, rank-based variant of the Mahalanobis distance is often used to limit the influence of outliers and of rare binary covariates; see Rosenbaum (2010, Chapter 8). This rank-based distance is not a metric on the Euclidean space of covariates, but it produces nonnegative numbers that satisfy \( \delta_{jk} \leq \delta_{ij} + \delta_{ik} \), so it yields a matching distance array.
In (3.1), the distance is a total, but when \( \rho \geq 2 \) or \( \kappa \geq 2 \), this total emphasizes the groups who will have more controls in the match. It is straightforward to alter given distances so that groups, rather than individuals within groups, receive equal emphasis. For any matching distance array \( \delta \), define \( \tilde{\delta}_{ij} = \delta_{ij}/\rho \), \( \tilde{\delta}_{ik} = \delta_{ik}/\kappa \) and \( \tilde{\delta}_{jk}'' = \delta_{jk}/(\rho\kappa) \). Because \( \delta \) is a distance array satisfying \( \delta_{jk}'' \leq \delta_{ij} + \delta_{ik}' \) and \( \rho \geq 1 \), \( \kappa \geq 1 \), it follows that \( \tilde{\delta} \) is a matching distance array satisfying \( \tilde{\delta}_{jk}'' \leq \tilde{\delta}_{ij} + \tilde{\delta}_{ik}' \). Computing (3.1) with \( \tilde{\delta} \) in place of \( \delta \) yields

\[
\widetilde{\delta}_B = \sum_{\ell=1}^{\rho} \tilde{\delta}_{i,j\ell} + \sum_{m=1}^{\kappa} \tilde{\delta}_{i,k,m} = \frac{1}{\rho} \sum_{\ell=1}^{\rho} \tilde{\delta}_{i,j\ell} + \frac{1}{\kappa} \sum_{m=1}^{\kappa} \tilde{\delta}_{i,k,m} = \frac{1}{\rho} \sum_{\ell=1}^{\rho} \tilde{\delta}_{i,j\ell} + \frac{1}{\kappa} \sum_{m=1}^{\kappa} \tilde{\delta}_{i,k,m},
\]

so the revised distance is an average rather than a total. The distinction between \( \delta \) and \( \tilde{\delta} \) will matter when \( \rho \) is large, because Step 2 of the approximation algorithm has one distance from \( I \) and \( \rho \) distances from \( J \) in each block distance to \( k \in K \). When \( \rho \) is large, \( \tilde{\delta} \) or some compromise between \( \delta \) and \( \tilde{\delta} \) may be more appropriate than \( \delta \) alone as a distance array.

Often, we wish to match exactly for a covariate, say belted or unbelted driver in Table 3. Because of the near-fine balance constraints in §3.3.2, matching exactly for one covariate with \( E \geq 2 \) categories and balancing another covariate with \( C \geq 2 \) categories is not the same as splitting the matching problem into \( E \) separate matching problems, say matching belted drivers, and separately matching unbelted drivers, because splitting attempts the more difficult task of balancing the \( E \times C \) joint categories. When either \( E \) or \( C \) or both are large, as is often true, near-fine balance of \( E \times C \) joint categories often entails tolerating larger deviations from fine balance than balancing \( C \) categories. When feasible, how can exact matching for \( E \) categories of one nominal variable be combined with near-fine balance for \( C \) categories of another?

To implement exact matching with near-fine balance constraints, start with any matching distance array \( \delta \), so \( \delta_{jk}'' \leq \delta_{ij} + \delta_{ik}' \). Define \( \tilde{\delta}_{ij} = \infty \) if \( i \) and \( j \) differ in their exact match category, otherwise \( \tilde{\delta}_{ij} = \delta_{ij} \), similarly define \( \tilde{\delta}_{ik} = \infty \) if \( i \) and \( k \) differ in exact match category, otherwise \( \tilde{\delta}_{ik} = \delta_{ik}' \), but define \( \tilde{\delta}_{jk}'' = \delta_{jk}'' \). Trivially, \( \tilde{\delta}_{jk}'' \leq \tilde{\delta}_{ij} + \tilde{\delta}_{ik}' \), so \( \tilde{\delta} \) is also a matching
distance array. Also trivially, if $i$ and $j$ are in the same exact category, and if $i$ and $k$ in the same exact category, then $j$ and $k$ are in the same exact category. It follows that a near-fine balance $(\rho, \kappa)$-design $B$ is exactly matched for the $E$-category variable if and only if $\delta_B < \infty$. Let $\overline{B}$ be a near fine $(\rho, \kappa)$-design whose distance $\overline{\delta_B}$ is minimal. If $\overline{\delta_B} = \infty$, then there is no near-fine $(\rho, \kappa)$-design $B$ that is exactly matched for the $E$-category variable. If $\overline{\delta_B} < \infty$ for the optimal design, then Proposition 1 implies that the near-fine $(\rho, \kappa)$-design $B$ produced by the approximation algorithm has finite total distance, $\overline{\delta_B} < \infty$, so that design is also exactly matched for the $E$ category variable. Additionally, if $\overline{\delta_B} < \infty$, then the bound, $\overline{\delta_B} \leq \overline{\delta_B} < \{1 + \max (\rho, \kappa)\} \overline{\delta_B}$ in Proposition 1 has avoided the infinite distances, and is referring to entries from the original matching distance array, $\delta$, for a match that is constrained to be both exact for the $E$ category variable and near-finely balanced for the $C$ category variable. In practice, distances that mismatch the $E$ exact categories are increased not to $\infty$, but rather increased by a large finite penalty, so infeasibility of near-fine balance joint with exact matching is recognized when such a penalized distance appears in the match; see the discussion of “almost exact” matching in Rosenbaum (2010, §9.2).

3.5. Distinguishing effects of side airbags and unmeasured biases

In every nonrandomized or observational study of treatment effects, observed associations may reflect effects caused by the treatment under study, or biases in who was treated, or a combination of the two. A strength of the design in §3.2 is that it provides several comparisons relevant to the effects of side airbags. These several comparisons may concur, strengthening evidence that associations are effects caused by side airbags, or they may clash suggesting that some or all associations are not effects of side airbags. A design that always encourages causal conclusions, a design that cannot suggest caution and restraint, is not a good design. An easy way to publish false causal conclusions is to decline to look for evidence that might reveal bias if present.

A study contains two or more evidence factors if it provides two or more tests of the null hypothesis of no treatment effect that would be (essentially) independent were there no
Figure 6: Injury severity in matched crashes grouped by the three eras, None, Optional, and All. The star in the boxplot and the vertical line in the barplot represent the mean.
Figure 7: Unmatched crashes in the optional era grouped by whether side airbags were present. The star in the boxplot and the vertical line in the barplot represent the mean. This interesting but unmatched comparison is not an evidence factor.
effect. A formal discussion of evidence factors involves technical issues that we do not present here; see Rosenbaum (2011; 2015, §3; 2017a; 2017b, §7). Recall that only 18% of studied vehicles in the optional period had side airbags.

Expressed more formally, our comparison contains several evidence factors, strict controls, and potential instruments. The “none” era is a control for the treated “all” era, because make and model determined the presence or absence of a side airbag in those eras, but those two eras are separated by several years. This study has two evidence factors: (i) all-versus-other-eras and (ii) optional-versus-none. The availability of side airbags during the optional era can serve as an instrument for the purchase of a side airbag when the optional era is compared to the other eras. Additionally, there is another comparison within the optional era, comparing cars with or without side airbags; however, these are typically comparisons between, rather than within, matched sets. We do not present these analyses in detail, because two graphs tell the whole story.

Figure 6 shows the injury severity sustained by the driver, broken down by era: “none”, “optional” or “all”. There is a substantial and statistically significant reduction in injury severity between “none” and “all”, but this is not plausibly an effect caused by side airbags, because the entire decline is already present in the “optional” era, when only 18% of owners had purchased vehicles with side airbags. Figure 7 is confined to the optional period, ignoring the matching, comparing vehicles with and without side airbags when owners had a choice about buying them. In Figure 7, the distribution of injuries looks similar with and without side airbags. In brief, the three-group design shows that some associations between side airbags and injuries are not plausibly effects caused by side airbags. With two rather than three groups, the same associations for two groups might mistakenly have been taken to be effects caused by side airbags.
3.6. Discussion

3.6.1. Summary

Observational studies often attempt to examine, possibly strengthen, a causal inference by making two comparisons instead of one treatment-versus-control comparison. The two comparisons may use two control groups, or two evidence factors, or combine a treated-control comparison with an instrument. Optimal construction of the relevant designs is essentially impossible; see Crama and Spieksma (1992, Theorem 1). We have proposed a polynomial time approximation algorithm that produces a near-fine \((\rho, \kappa)\)-design whose distance is not much greater than the unattainable optimal design.

3.6.2. Analyses with evidence factors or multiple control groups

The algorithm in Proposition 1 can be used to create two evidence factors, as in the example, or two control groups. The appropriate analysis depends upon the nature of the groups.

Two evidence factors provide two essentially independent tests of one null hypothesis of no treatment effect. These two independent tests and sensitivity analyses may be combined in a single analysis that combines independent \(P\)-values, such as the truncated product of Zaykin et al. (2002). The case \(\rho = 1, \kappa \geq 1\), is discussed in detail in Rosenbaum (2011), and it is illustrated in the `sensitivitymv` package in \texttt{R}; see Rosenbaum (2015b). The general case \(\rho \geq 1, \kappa \geq 1\), involves a stratified test (Rosenbaum 2018), rather than a matched test with multiple controls, and it may be implemented using the `senstrat` package in \texttt{R}; see also Rosenbaum (2017a) for elaboration.

Two control groups entail dependent tests. A simple strategy controls the family-wise error rate by testing several hypotheses in order, quitting when a null hypothesis is accepted. First, the treated group is compared to a combined control group, then to each control group separately, and finally comparisons attempt to show that the treated group differs more from both control groups than the control groups differ from each other; see Rosenbaum (2008) for specifics.
3.6.3. *More than three groups*

In most observational studies, finding two enlightening evidence factors or control groups is already a challenge, so we have focused on this situation, with a treated group and two comparison groups, or three groups in total. The situation with four or more groups is essentially parallel, and is implemented in the `approxmatch` package in R; see Crama and Spieksma (1992) for discussion of the simple case of groups of equal size without fine balance constraints. Step 2 of the algorithm is applied again, now matching blocks of the three group design to individuals from the fourth group, and so on. The approximation bound in Proposition 1 becomes worse because the triangle inequality is used again to bound some distances that were not optimized.
CHAPTER 4 : Evidence Factors in a Case-Control Study with Application to the
Effect of Flexible Sigmoidoscopy Screening on Colorectal Cancer

4.1. Introduction

4.1.1. Distal and proximal colon cancer and sigmoidoscopy screening

The U.S. Preventive Services Task Force (USPSTF) recommendations for colorectal cancer screening include flexible sigmoidoscopy every 5 years for men and women above 50 at average risk \cite{U.S. Preventive Services Task Force, 2016}. Yet, only 58% of adults aged 50–75 were up to date with the screening recommendations \cite{Joseph et al., 2016}. Is screening with sigmoidoscopy effective? Using a case-control study we aim to answer this question; more specifically we study the effect of screening by flexible sigmoidoscopy as per USPSTF recommendations on reducing mortality from colorectal cancer.

In case-control studies, patients with (cases) or without (controls) an outcome of interest are compared in terms of their exposure to treatment. Case-control studies are particularly useful for assessing treatment or exposure effects for rare outcomes. In a case-control study there is often a choice of how to define a case. In many setting, there are two (or more) ways to define a case, one being more “narrow” in that it is more likely to be caused by the exposure of interest if that exposure in fact has an effect and the other being “broad” in that it may have more heterogeneous causes.

Sigmoidoscopy can evaluate the lower or distal one-third of the colon for lesions; if abnormal, then a full colon evaluation with a colonoscopy is typically done for confirming the presence of cancer or precancerous polyps. The distal colon is the lower one-third part of the colon on the left side of the body, consisting of the descending colon, the sigmoid colon and the rectum; the proximal colon is the higher two-third of the colon. We consider broad cases to be all cases of colorectal cancer, and following Doubeni et al. \cite{Doubeni et al., 2018} and Selby et al. \cite{Selby et al., 1992}, we consider narrow cases to be cases where there are malignant polyps on the left side of the colon and rectum that are within the reach of the sigmoidoscope. We
expect that sigmoidoscopy screening, if it is effective, would only directly reduce the risk of
diagnosis or death from cancers in the distal colon (narrow cases) but would also indirectly
find or prevent some colorectal cancers in the proximal colon because abnormal findings in
the distal colon could trigger a colonoscopy. Is it possible to learn separate evidence about
the treatment effect when we have two or more definitions for a case? Before answering this
question in Section 4.1.3 we consider why one might want to construct separate evidence
and what we mean by separate evidence.

4.1.2. Evidence factors in an observational study

Unlike in a randomized trial, in a case-control study, as in any observational study, treat-
ment is not assigned to the subjects randomly. Therefore, a primary concern in a case-
control study is the potential for unmeasured confounders. In an observational study of
treatment effect, if there is concern that some unknown bias is the reason for the assess-
ment of significant treatment effect, we should consider if it possible to replicate the study
without repeating the bias (Cochran, 1965, Section 4.1).

For example, consider the effect of exposure to radiation on leukemia incidence. Radiolo-
gists, who are occupationally exposed to radiation, have been found to have a high incidence
of leukemia (Lewis, 1963). A replication of this observational study of the effect of radiation
that does not replicate the bias is a comparison of the leukemia risk in people living in Japan
near where the atomic bombs were dropped in Hiroshima and Nagasaki at the end of World
War II to people living further from the bombs (Bizzozero, Johnson, and Ciocco, 1966).
Radiologists may have higher rates of leukemia because they are more likely to diagnose it
and people living near the atomic bomb might have higher rates of leukemia because living
in an urban area may be a confounder for leukemia, but these are two different sources
of potential bias. The finding of higher rates of leukemia incidence in radiologists and in
people living near where the atomic bombs were dropped strengthens the evidence for a
causal effect since the skeptic would have assume two sources of bias rather than just one
(Rosenbaum, 2001).
In some studies, there may be two comparisons we can make within the same study that have different sources of bias, offering an opportunity for internal replication. When these comparisons are statistically independent or “nearly” independent (in the sense of (4.1)), the comparisons are called evidence factors (Rosenbaum, 2010a). A general perspective on evidence factors in an observational study is provided in Karmakar, French, and Small (2019a), which we briefly review here. Suppose two analyses are performed to test for the null hypothesis; the first analysis requires a set of assumptions $A_1$ and the second analysis requires a second set of assumptions $A_2$. Let $P_1$ and $P_2$ be the corresponding p-values from these two analyses. Then, to be evidence factors, we require the independence of the analyses in the following sense – under the null hypothesis, when both assumptions $A_1$ and $A_2$ hold for $(p_1, p_2) \in [0,1]^2$

$$
\Pr(P_1 \leq p_1, P_2 \leq p_2) \leq p_1 p_2. \quad (4.1)
$$

If the $P_1$ and $P_2$ were independent, then (4.1) would be an equality. The inequality in (4.1) means that the joint distribution of the p-values under the null hypothesis is stochastically bigger than that of two independent p-values under the null hypothesis, so that treating the two p-values as independent when using methods such as Fisher’s combination test (Fisher, 1932) to combine the two tests would be conservative – this is the “near independence” of the p-values we spoke of above. By asking for independence or near independence we ensure that we are learning two separate pieces of evidence rather than essentially one piece of evidence which would be the case if one uses two highly correlated tests such as a t-test and a Wilcoxon rank sum test (Rosenbaum, 2010a, 2011). We wish to avoid the mistake of the man who bought ‘several copies of the morning paper to assure himself that what it said was true’ (Wittgenstein, 1958, #265, quoted in Rosenbaum, 2010a). With the near independence from (4.1), if both analyses are significant, then we have two separate pieces of evidence against the null hypothesis of no treatment effect which are robust to violations of one of the two sets of assumptions $A_1$ or $A_2$. Both assumptions would have to be violated in order for there not to be evidence of a treatment effect. We give a formal definition of
An example of the use of evidence factors discussed in [Rosenbaum (2010a, 2015d)] that builds on a study of [Silber, et al. (2007)] is concerned with the effect of intensive chemotherapy for women with ovarian cancer. Chemotherapy for ovarian cancer can be administered by a medical oncologist (MO) or gynecological oncologist (GO); MOs tend to give higher doses of chemotherapy. In the same study, [Rosenbaum (2010a, 2015d)] considers the comparison of outcomes in pairs of women, one treated by an MO and one by a GO, who are matched for clinical stage and tumor grade, by a Wilcoxon signed rank test and also considers the comparison of the difference in outcomes between the MO patient and GO patient among pairs who have a higher difference in chemotherapy dose between the MO patient and GO patient vs. pairs with a lower difference in chemotherapy dose, by Kendall’s test. These tests are statistically independent under the null hypothesis of no treatment effect and also are likely to be biased by different factors – the comparison within pairs (the Wilcoxon signed rank test) might be biased by aspects of the way in which patients are referred to particular physicians while the comparison among pairs (Kendall’s test) might be biased by considerations that the physicians can see in their patients. [Rosenbaum (2010a)] found that both comparisons provide evidence that patients receiving higher amounts of chemotherapy have more weeks of toxicity after treatment. In contrast to a finding from just one comparison, the supportive findings from the two comparisons strengthens the evidence for more chemotherapy causally increasing toxicity because, in order for a skeptic to claim that more chemotherapy does not increase toxicity, the skeptic would have to assert two sources of bias rather than just one. Other examples of the use of evidence factors include [Zhang et al. (2011)] and [Zubizarreta et al. (2012)].

[Rosenbaum (2017)] provides a mathematical formulation for building evidence factors based on multiple treatment assignment mechanisms. Starting with a set of $n$ study units [Rosenbaum (2017)] showed how to construct evidence factors using the knit product of two subgroups of the symmetric group of size $n$. In this construction the assumptions made in the
two analyses state that the treatment is randomized in the corresponding subgroups.

Previous work has only considered constructing evidence factors based on different ways of assigning treatment. In this chapter, we develop novel evidence factors for case-control studies that use different definitions of a case. To the best of our knowledge, this is the first demonstration of using differences in outcomes to develop evidence factors.

4.1.3. Evidence factors in a case-control study with narrow and marginal cases

In a case-control study with narrow and broad cases, we expect that if the exposure has an effect and our theory that the narrow cases are more likely to be caused by the exposure than the more heterogeneous broad cases is correct and also there is no unmeasured confounding, then (a) the exposure should have a larger association with narrow cases than marginal cases, i.e. cases that are broad but not narrow and (b) the exposure should have an association with broad cases compared to controls. This is an elaborate theory of what a treatment effect, if there is an effect, is expected to look like. Elaborate theories, advocated by Sir Karl Popper and Sir Ronald Fisher, are an integral part of drawing causal conclusions from observational data (see Popper, 1959; Cochran, 1965, Section 5).

We compare the narrow cases to marginal cases to appraise association of pattern (a) in the elaborate theory and compare broad cases to controls to appraise association of pattern (b). These two comparisons could be biased differently. In the sigmoidoscopy study, unmeasured variables such as healthy lifestyle or greater compliance with medical treatment could be associated with screening. Some of these variables may be more associated with whether a person dies from any colorectal cancer or not (broad case vs. control) and some may be more associated with, among people who die from colorectal cancer, does the person die from a colorectal cancer on the distal colon or proximal colon (narrow case vs. marginal case). If we find evidence for both pattern (a) and pattern (b), this would require a skeptic to explain more types of bias than if we found one pattern alone; this point is developed formally in Section 4.6.
To test for patterns (a) and (b), we would like to use nearly independent test statistics in the sense of (4.1) so as to distinguish new evidence from the same evidence repeated twice. In other words, we would like to develop evidence factors associated with patterns (a) and (b). We develop a method for doing this in Section 4.4 and Section 4.5 proves that the test statistics developed are evidence factors. The data from the study is analyzed in Section 4.7, and in Section 4.8 a few other examples of case-control studies are discussed where multiple case definitions are used. Before developing our method, we discuss the data for the sigmoidoscopy study in Section 4.2 and provide notation in Section 4.3.

4.2. Sigmoidoscopy and colorectal cancer

Based on the reasoning of Section 4.1 we consider the effectiveness of screening sigmoidoscopy in relation to mortality from distal and proximal colon cancer. Specifically, we consider the comparison of distal cancer cases to proximal cancer cases and also the comparison of all cases (any colorectal cancer) to controls. In relation to sigmoidoscopy screening, distal cancer cases are narrow cases and proximal cancer cases are marginal cases. We will show later that these two comparisons constitute separate evidence (evidence factors). Throughout the chapter by sigmoidoscopy screening we mean specifically flexible sigmoidoscopy screening.

4.2.1. SCOLAR data

In a nested case-control study on members of Kaiser Permanente Northern California and Kaiser Permanente Southern California health-care systems study subjects were selected who were 55–90 years old between 2006 and 2012. Details of the study design are given in Goodman et al. (2015); Doubeni et al. (2018). A selected case unit would be a man or a woman who was 55–90 years old on the date of death with colorectal adenocarcinoma as the underlying cause of death. Using cancer diagnosis data and tumor characteristics 822 proximal and 886 distal cancer cases were identified. Each case patient was individually matched to controls on the reference date (which was the diagnosis date for each patient
Table 4: Balance on the covariates in the matched sets. Distal cancer cases are those who have been diagnosed to have died from cancer on the left colon or rectum, proximal cancer cases are from right colon cancer. For each covariate the mean is calculated within a matched set, then averaged over sets.

<table>
<thead>
<tr>
<th></th>
<th>Number of years enrolled before reference date</th>
<th>% from Center 1</th>
<th>% of female</th>
</tr>
</thead>
<tbody>
<tr>
<td>Controls</td>
<td>12</td>
<td>83</td>
<td>47</td>
</tr>
<tr>
<td>Distal cancer cases</td>
<td>12</td>
<td>83</td>
<td>46</td>
</tr>
<tr>
<td>Proximal cancer cases</td>
<td>12</td>
<td>84</td>
<td>47</td>
</tr>
</tbody>
</table>

who died of colorectal cancer), gender, the duration of health plan prior to diagnosis and the health-care site. In this process 3635 controls were included.

Thus in our design, there are 822 narrow cases and 886 marginal cases. To facilitate the comparison of narrow cases to marginal cases we pair matched narrow (distal cancer) cases to marginal (proximal cancer) cases using the `optmatch` package in R which uses methods of Hansen and Klopfer (2006). The matching algorithm used a weighted sum of rank based Mahalanobis distance and absolute distance of estimated logit propensity scores. It also fine balanced on gender. By pair matching the narrow and marginal cases, we obtained 822 matched sets consisting of one narrow case, one marginal cases and the controls associated with these cases and $886 - 822 = 64$ matched sets consisting of one marginal case and the controls associated with this case. Table 4 shows the covariate balance of the matched sets. Figure 8 further shows the distribution of the diagnosis year of the colorectal cancer patients. Gender, reference date and enrollment source are well balanced between the narrow cases, marginal cases and controls over the matched sets.

Although the match controls well for gender, reference date and enrollment source, there could be unmeasured confounders. For example, lack of physical activity is a known risk factor of colorectal cancer incidence and people who are less active also may be less likely to get screened (Eldridge et al., 2013). Because we are not able to match on or adjust for physical activity in our analysis, the comparison of all colorectal cancer cases to controls
may be biased. Family history of cancer screening is another likely unmeasured confounder in this analysis. The comparison of sigmoidoscopy screening in proximal vs. distal cancers may also be biased by unmeasured confounding. There are potential biological differences between proximal and distal colon cancers such that variables such as diet (e.g., use of the Mediterranean diet) may be differentially associated with proximal and distal colon cancer (Doubeni et al., 2012; Missiaglia et al., 2014). Such diet choices may be associated with screening. If we find that sigmoidoscopy screening is associated with reduced mortality from colorectal cancer when comparing all cases to controls and with reduced mortality from proximal vs. distal cancer cases when comparing proximal to distal cases, then, in order for these associations to arise purely from bias and not at all from a causal effect of sigmoidoscopy screening on reducing cancer, there would need to be unmeasured confounders in both comparisons rather than just one comparison. In Section 4.6, we show that even if the unmeasured confounders for the two comparisons overlap but have different relative magnitudes, the evidence is strengthened by finding significant associations in both comparisons.

Figure 8: Reference date of the colorectal cancer cases and controls in the matched sets.
As suggested earlier we shall assess the effect of sigmoidoscopy screening by comparing the prevalence of screening between all colorectal cancer cases and controls and also by comparing the prevalence between the distal cancer cases and proximal cancer cases. Results of this analysis will be discussed in Section 4.7. We first present the methodology.

4.3. Notation and review: case-control studies

Let observational units be denoted by indices $l = 1, \ldots, L$. We use the binary variable $Z_l$ to denote whether unit $l$ was exposed to treatment ($Z_l = 1$) or spared from being exposed ($Z_l = 0$). Under the potential response model suppose unit $l$ if exposed would have response $r_{Tl}$ and if spared exposure would have response $r_{Cl}$. Thus the observed response for unit $l$ is $R_l = Z_l r_{Tl} + (1 - Z_l) r_{Cl}$. Consequently, we cannot observe $r_{Tl}$ and $r_{Cl}$ simultaneously for one unit (Neyman, 1923; Rubin, 1974). Now let $x_l$ denote the observed pre-treatment covariates, i.e., covariates recorded in the study that can potentially affect the treatment assignment and the response. The unobserved confounders are summarized by an unobserved number $u_l$ for unit $l$ scaled to be valued in $[0, 1]$ (Rosenbaum, 1991). We gather population parameters to write the set $F = \{ (r_{Tl}, r_{Cl}, x_l, u_l) : l = 1, \ldots, L \}$. The hypothesis we are interested in studying is Fisher’s sharp null hypothesis of no treatment effect

$$H_0: r_{Tl} = r_{Cl}, \ l = 1, \ldots, L.$$

A case definition is a function $k(\cdot)$ which labels each unit as a case or a control or neither based on the observed response. A case definition would identify a subset of the units as cases and a separate subset as controls.

For a given case definition, a test for the hypothesis $H_0$ can be carried out by matching as follows. Create $S$ strata labeled $s = 1, \ldots, S$ where each stratum consists of a total of $t_s$ units with some case units and the rest control units (say $c_s$) which are similar with respect to the observed covariates ($x_l$’s). Now let $Y_s$ denote the total number of exposed case units in stratum $s$. A positive linear combination $T = \sum_{s=1}^{S} d_s Y_s$ can be taken as a test statistic
for testing the hypothesis $H_0$. When all $d_s = 1$ the statistic $T$ is exactly the total number of exposed cases, which is the Mantel-Haenszel test statistic.

We assume that the treatment assignments for distinct units are independent. We consider the following model for treatment assignment

$$
\Pr(Z_l = 1 \mid \mathcal{F}) = \frac{\exp\{\lambda(x_l) + \gamma u_l\}}{1 + \exp\{\lambda(x_l) + \gamma u_l\}},
$$

(4.2)

where $\lambda(\cdot)$ is an unknown function and $\gamma \geq 0$ is an unknown parameter. Since $0 \leq u_l \leq 1$, for two units $l$ and $l'$ ($l \neq l'$) with the same observed covariates, $x_l = x_{l'}$, under this model, their odds of exposure can vary at most by a factor of $\Gamma := \log(\gamma)$. Model (4.2) is equivalent to writing

$$
\max_{1 \leq l, l' \leq L} \left\{ \frac{\Pr(Z_l = 1 \mid \mathcal{F})/\Pr(Z_l = 0 \mid \mathcal{F})}{\Pr(Z_{l'} = 1 \mid \mathcal{F})/\Pr(Z_{l'} = 0 \mid \mathcal{F})} : x_l = x_{l'} \right\} \leq \Gamma.
$$

(4.3)

The fact that (4.2) implies (4.3) is obvious, the proof of the reverse implication constructs a set of $u_l$ from the odds of exposure ([Rosenbaum, 2002, Section 4.4.4]). The parameter $\Gamma(\geq 1)$ is the hidden bias level. Thus, when $\Gamma = 1$, there is no unmeasured confounder and there is no bias in treatment assignment after controlling for observed covariates. As $\Gamma$ increases, this model allows more and more bias in treatment assignment. For example, when $\Gamma = 2$, due to the presence of unmeasured confounders, it might be possible that for individuals who are the same in their observed covariates, one individual has twice the odds of getting assigned treatment as the other or vice versa.

Let $e_s$ be the number of exposed units in stratum $s$. Then under model (4.3), we can bound the tail probability of $T$ under $H_0$ asymptotically

$$
\Pr(T \geq k \mid \{t_s\}, \{c_s\}, \{e_s\}, \mathcal{F}) \leq 1 - \Phi \left( \frac{k - \sum d_s(t_s - c_s)p_s}{\sqrt{\sum d_s^2(t_s - c_s)p_s(1 - p_s)}} \right),
$$

(4.4)

where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution and $\bar{p}_s = \Gamma e_s/(\Gamma e_s + (t_s - e_s))$ ([Small et al., 2013]). This tail bound is sharp in that it is attained

Therefore, given a case-control study, after constructing a satisfactory stratum structure, when the hidden bias level is at most \( \Gamma \), i.e., (4.3) holds, (4.4) can be used to get an upper bound for the p-value of testing the hypothesis \( H_0 \). If this value is less than \( \alpha \), the significance level, then we have evidence to reject the null hypothesis as long as the hidden bias is at most \( \Gamma \). A sensitivity analysis asks how much bias in the treatment assignment must be present so that the observed association can be explained just from bias under \( H_0 \). Thus a sensitivity analysis looks for \( \Gamma \) values where the upper bound of (4.4) is larger than \( \alpha \).

4.4. Two case definitions and two comparisons

Following our discussion in Section 4.1.3 consider a design with availability of two case definitions, one narrow and one broad. A case unit according to a narrow case definition is also a case unit in a broad case definition. We label a unit as marginal case unit if it is not a case in a narrow case definition but is a case in broad case definition. The study units which are non-cases in broad case definition are thus also non-cases in the narrow case definition, and are labeled as controls.

4.4.1. Matched strata for the comparisons

In Section 4.3 we have discussed a matching procedure for a comparison between cases and controls to test for \( H_0 \) when there is one standard case definition. Now as we have two case definitions, a similar matching argument can still be used with appropriate modifications. We discuss the details here.

Suppose the matching procedure creates \( S \) strata of all three types of units: narrow cases, marginal cases and controls where units in a stratum are similar in their observed covariates. Let a generic stratum labeled \( s \) have \( n_s \) narrow cases, \( m_s \) marginal cases, thus a total of \( b_s = n_s + m_s \) broad cases and \( c_s \) controls. In a cohort of \( L \) units a narrow case definition might have a much smaller number of cases than a broad case definition. In such situations
some of the stratum \((s)\) may only have marginal cases and controls, resulting in \(n_s = 0\), which is allowed in our notation. But each stratum must consist of at least two different labels of units. Let the letters \(n\), \(m\), \(b\) or \(c\) for denoting that the unit is a narrow case, a marginal case, a broad case or a control respectively. For example, \(Z_{n(si)}\) denotes the exposure \((0\ or\ 1)\) for the \(i\)th narrow case in the stratum \(s\) \((s\ in\ 1, 2, \ldots, S)\). The index \(i\) runs in \([n_s]\) (we use the notation \([k]\) to denote the set \(\{1, \ldots, k\}\) if \(k\) is a positive integer or empty set \(\{\}\) otherwise). Similarly \(x_{c(si)}\) denotes the observed covariate for the \(i\)-th control in stratum \(s\). \(R_{m(si)}, r_{Cn(si)}, u_{c(si)}\) etc. have similar meanings.

At this point we can quantify the evidence against \(H_0\) by calculating the \(p\)-values from the two comparisons of narrow cases versus marginal cases and broad cases versus controls. We focus on the linear statistics of the number of exposed narrow cases and broad cases respectively for these two comparisons. Let \(Y_{n(si)}\) and \(Y_b(s)\) for stratum labeled \(s\) denote the number of exposed narrow cases and the number of exposed broad cases. Notice that,

\[
Y_b(s) = \sum_{i \in [n_s]} Z_{n(si)} + \sum_{i \in [m_s]} Z_{m(si)} = Y_{n(s)} + Y_{m(s)}.
\]

Two test statistics for these two comparisons can be written as

\[
T_{nm} = \sum_{s=1}^{S} d_{nm(s)} Y_{n(s)} \quad \text{and} \quad T_{bc} = \sum_{s=1}^{S} d_{bc(s)} Y_{b(s)},
\]

where \(d_{nm(s)}\) and \(d_{bc(s)}\) are non-negative constants given \(\mathcal{F}\).

Under assumption (4.3) about treatment assignment distribution, we can get bounds on the \(p\)-values for \(T_{nm}\) and \(T_{bc}\). But there are a few subtleties here that are important to point out.

First, a \(p\)-value for \(T_{nm}\) should only be based on information from the narrow cases and marginal cases. In other words the \(p\)-value \(P_{nm}\) is computed based on the tail distribution

\[
\Pr\left(T_{nm} \geq k \mid \{b_s\}, \{m_s\}, \sum_{i \in [n_s]} Z_{n(si)} + \sum_{i \in [m_s]} Z_{m(si)}, \mathcal{F}_b\right), \quad (4.5)
\]

where \(\mathcal{F}_b\) is the subset of \(\mathcal{F}\) restricted to the broad cases. In equation (4.4), \(t_s\) was used
instead of $b_s$, $c_s$ was used instead of $m_s$ and the sum above replaces $e_s$. Similarly, the p-value $P_{bc}$ is computed based on the tail distribution

$$
\Pr\left( T_{bc} \geq k \mid \{b_s + c_s\}, \{c_s\}, \sum_{i \in [n_s]} Z_{n\{si\}} + \sum_{i \in [m_s]} Z_{m\{si\}} + \sum_{i \in [c_s]} Z_{c\{si\}}, F \right). \tag{4.6}
$$

Thus in technical terms, $P_{nm}$ and $P_{bc}$ are measurable with respect to different sigma fields.

Second, in assumption (4.3) the sensitivity parameter $\Gamma$ bounds the odds ratio of treatment assignment for all the units stratified on their observed covariates. But as noted in the previous subsection, unmeasured confounders are likely to affect the two comparisons in different ways (see also Section 4.6). Therefore, while considering narrow versus marginal comparison we should relax this assumption only to the broad cases since these are the only ones contributing to $T_{nm}$. Hence, we distinguish the effect of unmeasured covariates for the two comparisons by using two sensitivity parameters $\Gamma_{nm}$ and $\Gamma_{bc}$ for the narrow versus marginal and broad versus control comparisons respectively. Then $\Gamma_{nm}$ measures the bias in treatment assignment among all the case units which are similar in their observed covariates and $\Gamma_{bc}$ measures the bias in treatment assignment among all case and control units which are similar in their observed covariates.

Therefore, the comparison of narrow versus marginal cases would compute the upper bound on the p-value for $T_{nm}$ based on the tail distribution (4.5) for sensitivity parameter $\Gamma_{nm}$ and the broad cases versus controls comparison would compute the upper bound on the p-value for $T_{bc}$ based on the tail distribution (4.6) for sensitivity parameter $\Gamma_{bc}$. We denote them by $P_{nm,\Gamma_{nm}}$ and $P_{bc,\Gamma_{bc}}$ respectively and when $\Gamma_{nm} = \Gamma_{bc} = 1$ we simply write $P_{nm}$ and $P_{bc}$ for $P_{nm,1}$ and $P_{bc,1}$ respectively.

4.4.2. Two sensitivity parameters and their amplification

In a sensitivity analysis the sensitivity parameters $\Gamma_{nm}$ and $\Gamma_{bc}$ would be used to get the max p-values $P_{nm,\Gamma_{nm}}$ and $P_{bc,\Gamma_{bc}}$. How does an $\Gamma_{nm}$ bias relate to the influence of the
unmeasured confounding on the exposure to treatment of an unit and the influence of the unmeasured confounding on the narrow to marginal case status of the unit? The sensitivity analysis model (4.2) conditions on the information set \( \mathcal{F} \), which includes the potential outcomes of the units. The maximum p-value calculated under this model is achieved when there is a near perfect relationship between the case definition and the unmeasured confounders. We discuss here that this model can be interpreted differently, “amplified,” to be a model that limits the relationship between the case definition and the unmeasured confounders as well as the relationship between the exposure and the unmeasured confounders (Gastwirth, Krieger, and Rosenbaum, 1998; Rosenbaum and Silber, 2009).

Let the confounding variable in the broad cases to controls comparison be \( u_1 \) and the confounding variable in narrow to marginal comparison be \( u_2 \). Consider now the set \( \mathcal{C} = \{(x_i, u_{1l}, u_{2l}) : l = 1, \ldots, L\} \). As before, \( 0 \leq u_{1l} \leq 1 \) and \( 0 \leq u_{2l} \leq 1 \). A principal stratification on the set \( \mathcal{C} \) does not condition on the potential outcomes.

Consider two units \( i_1 \) and \( i_2 \) with the same observed covariates. We model the relationship between the unmeasured confounding and the treatment assignment with a parameter \( \lambda \), for \( z_{i1} + z_{i2} = 1 \), as

\[
\Pr(Z_{i1} = z_{i1}, Z_{i2} = z_{i2} \mid \mathcal{C}, x_{i1} = x_{i2}, Z_{i1} + Z_{i2} = 1) = \frac{\exp\{\lambda(z_{i1}w_{1l} + z_{i2}w_{2l})\}}{\exp(\lambda w_{1l}) + \exp(\lambda w_{2l})}, \tag{4.7}
\]

where \( w_l = \xi_1 u_{1l} + \xi_2 u_{2l} \), for \( l = 1, \ldots, L; \xi_1, \xi_2 \geq 0, \xi_1 + \xi_2 = 1 \). \tag{4.8}

If \( \lambda = 0 \) the probability is \( 1/2 \) and the confounders have no effect. A larger value of \( \lambda \) indicates a larger influence of the unmeasured confounders on the treatment assignment. Equation (4.8) in itself is not a new assumption. Any number \( w_l \), taking value in \([0,1]\), can be rewritten as \( w_l = \xi_1 u_{1l} + \xi_2 u_{2l} \), for \( \xi_1, \xi_2 \geq 0, \xi_1 + \xi_2 = 1 \) and \( 0 \leq u_{1l}, u_{2l} \leq 1 \), and vice versa. Hence, this model is similar in spirit to model (4.2) except that the principal conditioning now changes from \( \mathcal{F} \) to \( \mathcal{C} \).

Next we model the relationship of the unmeasured confounding and the case status. Let us
denote for unit \( l \), when not exposed to the treatment, by the indicator variable \( k_{Cl}^b \) whether the unit is a case and by \( k_{Cl}^n \) whether the unit is a narrow case. Thus, \( k_{Cl}^b = 1 \) if the \( l \)th unit is a case, either narrow or marginal, when not exposed to the treatment and \( k_{Cl}^b = 0 \) if the unit is a control when not exposed to the treatment. Similarly, \( k_{Cl}^b = 1 \) if the \( l \)th unit is a narrow case when not exposed to the treatment and \( k_{Cl}^b = 0 \) otherwise. It might be helpful to think of \( k_{Cl}^b \) and \( k_{Cl}^n \) as being determined by \( r_{CI} \). For two units \( i_1 \) and \( i_2 \) with similar observed covariates, the following model relates the case label with the confounders:

\[
\begin{align*}
\Pr(k_{Cl1}^b = 1, k_{Cl2}^b = 0 \mid C, x_{i1} = x_{i2}) &= \exp\{\delta_{bc}(u_{1,i1} - u_{1,i2})\}; \\
\Pr(k_{Cl1}^n = 1, k_{Cl2}^n = 0 \mid C, x_{i1} = x_{i2}, k_{Cl1}^b = k_{Cl2}^b = 1) &= \exp\{\delta_{nm}(u_{2,i1} - u_{2,i2})\}.
\end{align*}
\]

The level of bias from unmeasured confounding \( u_1 \) in being a broad case is \( \delta_{bc} \), and the level of bias from unmeasured confounding \( u_2 \) in being a narrow case over a marginal case is \( \delta_{nm} \). The larger the value of these parameters the higher the influence of the unmeasured confounding.

How do \( \Lambda, \delta_{bc} \) and \( \delta_{nm} \) relate to the sensitivity parameters \( \Gamma_{bc} \) and \( \Gamma_{nm} \)? Proposition 1 of Rosenbaum and Silber (2009) provides the correspondence. Let \( \Lambda = \exp(\lambda) \), \( \Delta_{bc} = \exp(\delta_{bc}) \) and \( \Delta_{nm} = \exp(\delta_{nm}) \). Then \( \Gamma_{bc} = (\Delta_{bc}\Lambda + 1)/(\Delta_{bc} + \Lambda) \) and \( \Gamma_{bc} = (\Delta_{nm}\Lambda + 1)/(\Delta_{nm} + \Lambda) \).

These formulas allow one to interpret the result of a sensitivity analysis either using the sensitivity parameters \( \Gamma_{bc} \) and \( \Gamma_{nm} \) or, under model (4.7)–(4.10), using parameters \( \lambda, \delta_{bc} \) and \( \delta_{nm} \). For example, \( \Gamma_{nm} = 2, \Gamma_{bc} = 2.5 \) corresponds to \( \Lambda = 4, \Delta_{nm} = 3.5 \) and \( \Delta_{bc} = 8 \); \( \Gamma_{nm} = 3, \Gamma_{bc} = 2 \) corresponds to \( \Lambda = 5, \Delta_{nm} = 7 \) and \( \Delta_{bc} = 3 \) and so on.

4.5. Evidence factors

This section aims to establish that the two comparisons discussed in Section 4.4.1 explore different aspects of the study design and give separate evidence, and thus are evidence factors. The idea of evidence factors was first formalized by Rosenbaum (2010a) and extended for studies with multiple treatment assignment mechanisms in Rosenbaum (2011, 2017). As
discussed in Section 4.1.2, Karmakar, French, and Small (2019a) provide a general formulation of evidence factors in observational study designs. We start this section by stating this definition. Readers interested in the results of the SCOLAR data analysis can skip this technical discussion and go directly to Section 4.5.1 and 4.7.

**Definition 6** A set $D$ is called a decreasing set if for any pair $(x, y)$ with $x \leq y$, if $y \in D$ then $x \in D$. For two random vectors $X$ and $Y$ we say that $X$ is stochastically larger than $Y$ if

$$Pr(X \in D) \leq Pr(Y \in D),$$

(4.11)

for all nondecreasing sets $D$. If $X$ is stochastically larger than $Y$ we write $X \succ Y$.

**Definition 7** For any pair of bias levels $(\Gamma_{nm}, \Gamma_{bc})$, $(P_{nm, \Gamma_{nm}}, P_{bc, \Gamma_{bc}})$ are evidence factors for testing $H_0$ if, $(P_{nm, \Gamma_{nm}}, P_{bc, \Gamma_{bc}}) \succ (U_1, U_2)$ under the bias levels $\Gamma_{nm}$, $\Gamma_{bc}$ and under $H_0$, for two independent Unif[0,1] random variables $U_1$ and $U_2$.

The main result of this section is the following theorem which says that according to the above definition, $(P_{nm, \Gamma_{nm}}, P_{bc, \Gamma_{bc}})$ form evidence factors.

**Theorem 8** Under $H_0$ and for bias levels $\Gamma_{nm}$ and $\Gamma_{bc}$ we have $(P_{nm, \Gamma_{nm}}, P_{bc, \Gamma_{bc}}) \succ (U_1, U_2)$ for two independent Unif[0,1] random variables $U_1$ and $U_2$.

The rest of the section is dedicated to proving this theorem using a few lemmas. The proof of all the lemmas are given in the appendix.

To slightly simplify our notation in what follows, for two random vectors $X$ and $Y$ we write $[X \mid Y]$ to denote the conditional distribution of $X$ given $Y$. Since we are dealing with discrete spaces, $[X \mid Y]$ is a real valued measurable function of $X$ and $Y$.

The following is one of the main lemmas needed to prove Theorem 8.
Lemma 3 There exists functions $f_{nm}$ and $f_{bc}$ on appropriate domains such that

$$P_{nm} \Gamma_{nm} = f_{nm}\left(\left\{Z_{n\{si\}}, i \in [n_s]; \sum_{i \in [n_s]} Z_{n\{si\}} + \sum_{i \in [m_s]} Z_{m\{si\}} \mid s \in [S]\right\}\right),$$

and

$$P_{bc} \Gamma_{bc} = f_{bc}\left(\left\{Z_{c\{si\}}, i \in [c_s]; \sum_{i \in [n_s]} Z_{n\{si\}} + \sum_{i \in [m_s]} Z_{m\{si\}} \mid s \in [S]\right\}\right).$$

Following Definition 6, let us use the notation $X \succ D$ for a random variable $X$ and a probability distribution $D$ to say that $X$ is stochastically larger than $D$ or $\Pr(X \leq x) \leq \Pr(Y \leq x \mid Y \sim D)$ for all $x \in \mathbb{R}$.

Lemma 4 Under $H_0$, we have the following

(i) $[P_{nm} \Gamma_{nm} \mid \{\sum_{i \in [n_s]} Z_{n\{si\}} + \sum_{i \in [m_s]} Z_{m\{si\}}\}; \mathcal{F}, \{n_s\}] \succ \text{Unif}[0, 1]$.

(ii) $[P_{bc} \Gamma_{bc} \mid \{\sum_{i \in [n_s]} Z_{n\{si\}} + \sum_{i \in [m_s]} Z_{m\{si\}} + \sum_{i \in [c_s]} Z_{c\{si\}}\}; \mathcal{F}, \{b_s + c_s\}] \succ \text{Unif}[0, 1]$.

(iii) $P_{nm} \Gamma_{nm} \succ \text{Unif}[0, 1]$.

(iv) $P_{bc} \Gamma_{bc} \succ \text{Unif}[0, 1]$.

The following lemma relies on the assumption of no interference in treatment assignment among the units, which is to say $Z_l$ and $Z_{l'}$ are independently distributed for two distinct units $l$ and $l'$.

Lemma 5 Under $H_0$

$$[P_{nm} \Gamma_{nm} \mid \{Z_{c\{si\}}, i \in [c_s]\}; \sum_{i \in [n_s]} Z_{n\{si\}} + \sum_{i \in [m_s]} Z_{m\{si\}}] \succ \text{Unif}[0, 1].$$

Lemma 6 Suppose two random variables $P_1$ and $P_2$ satisfy

$C1$ random variable $P_1$ is a function of random quantity $V_1$,

$C2$ $[P_2|V_1] \succ \text{Unif}[0, 1]$,
then for \( 0 \leq q \leq 1 \), \( \Pr(P_2 \leq q \mid P_1) \leq q \), i.e., \( [P_2 \mid P_1] \approx \text{Unif}[0,1] \).

Now we have all the necessary facts to prove Theorem 8.

**Proof.** (of Theorem 8) In Lemma 6 take \( P_1 = P_{bc,\Gamma_{bc}} \), \( P_2 = P_{nm,\Gamma_{nm}} \) with \( V_1 = \{Z_{c\{si\}}, i \in [c_s]\}; \sum_{i \in [n_s]} Z_{n\{si\}} + \sum_{i \in [m_s]} Z_{m\{si\}} \}. \) Then, by Lemma 3, condition C1 is satisfied and condition C2 is proved in Lemma 5. Thus by Lemma 6, \( [P_{nm,\Gamma_{nm}} \mid P_{bc,\Gamma_{bc}}] \approx \text{Unif}[0,1] \).

Let \( U_1 \) and \( U_2 \) be two independent uniformly distributed random variables on \([0,1]\). We use the theory of Shaked and Shanthikumar (2007) §6B. \((U_1, U_2)\) being an independent pair is a conditionally increasing in sequence (CIS). Then combining this with the facts that \( P_{bc,\Gamma_{bc}} \approx \text{Unif}[0,1] \) (by Lemma 4) and \( [P_{nm,\Gamma_{nm}} \mid P_{bc,\Gamma_{bc}}] \approx \text{Unif}[0,1] \), Theorem 6.B.4 of Shaked and Shanthikumar (2007) finally gives us

\[
(P_{nm,\Gamma_{nm}}, P_{bc,\Gamma_{bc}}) \approx (U_1, U_2).
\]

Thus the proof is complete.

4.5.1. **Combining evidence**

In words, Theorem 8 says that the combined information from the two evidence \( P_{nm,\Gamma_{nm}} \) and \( P_{bc,\Gamma_{bc}} \) carries as much evidence as two independent evidence. This allows us to combine these two pieces of evidence and provide a total evidence against the hypothesis under both the comparisons. Karmakar et al. (2019a) discusses different methods for combining evidence. Any method of combining p-values that is monotone in both of the p-values can be used, e.g. Fisher’s combination method (Fisher, 1932), the mean of the normal transformation (Liptak, 1958), the truncated product method of combining (Hsu et al., 2013; Zaykin et al., 2002). Also see Becker (1994).

Fisher’s method computes the joint evidence as the tail probability of \( \chi^2 \) distribution over

\[ -2 \log(P_{nm,\Gamma_{nm}} \cdot P_{bc,\Gamma_{bc}}). \]

In the scenario of sensitivity analysis, since we only consider largest possible p-values for a given value of hidden bias level, the truncated product method, which
weights the evidence by the strength of the evidence is often preferred. For a given \( \alpha \), the
combined evidence using the truncated product method is given by
\[ F_W(E_{\Lambda, \Gamma_{nm}, \Gamma_{bc}}), \]
where
\[ E_{\Lambda, \Gamma_{nm}, \Gamma_{bc}} = \mathbb{1}_{P_{nm, \Gamma_{nm}} \leq \alpha} \log(P_{nm, \Gamma_{nm}}) + \mathbb{1}_{P_{bc, \Gamma_{bc}} \leq \alpha} \log(P_{bc, \Gamma_{bc}}), \]
and (4.12)
\[ F_W(w) = 2\alpha(1 - \alpha)[1 - F_{\text{Exp}}(1)(-\log(\frac{w}{\alpha^2}))] + \alpha^2[1 - F_{\text{Gamma}(2,1)}(-\log(\frac{w}{\alpha^2}))]. \]
The advised choice of \( \alpha \) is 0.20 ([Hsu et al., 2013; Zaykin et al., 2002]).

We conducted a simulation study to compare the powers of Fisher’s method and the truncated product method in the setting of our problem. The simulation scenario considered here is based on the case-control study structure. We are going to look at the favorable situation where there are no unmeasured confounders with treatment effect. Then for varied treatment effect sizes, we compare the power of the two combining methods for different values of \( (\Gamma_{nm}, \Gamma_{bc}) \).

We consider a population where the chance of exposure is 1/3. Thus for a unit \( l \), \( \Pr(Z_l = 1) = 1/3 \). The treatment effect is denoted by \( \beta \). We consider a univariate response and two types of response distributions in the population. The two types of distributions when spared exposure are a normal distribution with mean 0 and variance 1 and a t-distribution normalized to have variance 1. Therefore, if a unit \( l \) is exposed to treatment then the response is a sample from \( N(\beta, 1) \) (or \( \beta + t_3/\sqrt{3} \)) and if not exposed then the response is a sample from \( N(0, 1) \) (or \( t_3/\sqrt{3} \)). The case-definition for each of the scenarios is taken such that if the treatment effect was 0.5 then 20% of the population would be broad cases.

Thus, in the setting where the response is from normal distribution, the response of more than the 0.8 quantile of the mixture distribution \( 1/3N(\beta, 1) + 2/3N(0, 1) \) would be labeled a broad case. In our simulation we sample 2,000 broad cases and largest half of them are labeled as narrow cases. Then we sample 2,000 controls. In both comparisons of narrow cases versus marginal cases and broad cases versus controls we consider paired stratum, i.e., \( n_s = m_s = 1, c_s = 2 \).
Table 5: Simulated power of a sensitivity analysis of combined evidence in a case control study, where there is no unmeasured confounder and \( \Pr(Z_l = 1) = 1/3 \). The response is simulated from \( N(\beta, 1) \) if \( Z_l = 1 \) and \( N(0, 1) \) if \( Z_l = 0 \). There are 1,000 narrow cases and 1,000 marginal cases with 2,000 controls. Based on 10,000 iterations. Fisher = Fisher’s combination method, \( tP = \) truncated product method with \( \tilde{\alpha} = 0.20 \).

<table>
<thead>
<tr>
<th>( \Gamma_{nm} )</th>
<th>( \Gamma_{bc} )</th>
<th>( \beta = 0 )</th>
<th>( \beta = 0.2 )</th>
<th>( \beta = 0.4 )</th>
<th>( \beta = 0.6 )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Fisher tP</td>
<td>Fisher tP</td>
<td>Fisher tP</td>
<td>Fisher tP</td>
<td>Fisher tP</td>
</tr>
<tr>
<td>1</td>
<td>0.0491 0.0463</td>
<td>0.9999 1</td>
<td>1 1</td>
<td>1 1</td>
<td></td>
</tr>
<tr>
<td>1.5</td>
<td>0.0061 0.0102</td>
<td>0.2525 0.257</td>
<td>1 1</td>
<td>1 1</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0.0061 0.0102</td>
<td>0.1758 0.2236</td>
<td>0.8671 0.8564</td>
<td>1 1</td>
<td></td>
</tr>
<tr>
<td>2.5</td>
<td>0.0061 0.0102</td>
<td>0.1758 0.2236</td>
<td>0.7478 0.7971</td>
<td>0.9993 0.9991</td>
<td></td>
</tr>
</tbody>
</table>

| 1.25 | 1.25 | 0 0 | 0.4807 0.507 | 1 1 | 1 1 | 1 1 |
| 2    | 0 0  | 0.0005 0.0011 | 0.1552 0.1544 | 0.9999 | 1 |
| 2.75 | 0 0  | 0.0005 0.0011 | 0.0318 0.0459 | 0.6865 | 0.6638 |
| 3.5  | 0 0  | 0.0005 0.0011 | 0.0318 0.0459 | 0.6865 | 0.6638 |

| 1.5 | 1.5 | 0 0 | 0.0023 0.0033 | 0.9921 0.9936 | 1 1 |
| 2   | 0 0 | 0 0 | 0.0001 0.0001 | 0.5444 | 0.5243 |
| 3.5 | 0 0 | 0 0 | 0.0001 0.0001 | 0.0135 | 0.0205 |

| 1.75 | 1.75 | 0 0 | 0 0 | 0.5115 0.5784 | 1 1 |
| 2    | 0 0 | 0 0 | 0 0 | 0.0217 0.0342 | 0.9983 | 0.9989 |
| 3.25 | 0 0 | 0 0 | 0 0 | 0 0 | 0.0001 | 0.0002 |

| 2   | 2   | 0 0 | 0 0 | 0.0215 0.0342 | 0.9983 | 0.9989 |
| 2.5 | 0 0 | 0 0 | 0 0 | 0 0 | 0.3569 | 0.4188 |
| 3   | 0 0 | 0 0 | 0 0 | 0 0 | 0.0001 | 0.0002 |
| 3.5 | 0 0 | 0 0 | 0 0 | 0 0 | 0 0 | 0 0 |

| 2.25 | 2.25 | 0 0 | 0 0 | 0 0 | 0 0 | 0.8795 | 0.9117 |
| 2.5  | 0 0 | 0 0 | 0 0 | 0 0 | 0 0 | 0.3542 | 0.4187 |
| 3    | 0 0 | 0 0 | 0 0 | 0 0 | 0 0 | 0.0009 | 0.0022 |

Table 5 and 6 report the simulated power for the two combining methods. The simulated power is based on 10,000 iterations with level of significance \( \alpha = 0.05 \). Except for very few situations in Table 5, the truncated product method has better simulated power than Fisher’s combining method. The truncated product method seem to be less sensitive as we increase \( \Gamma_{nm} \) and \( \Gamma_{bc} \). Fisher’s method has slightly better simulated power in a few situations in the normal response model for moderate values of \( (\Gamma_{nm}, \Gamma_{bc}) \) when there is a large treatment effect (\( \beta = 0.6 \)). After considering these simulation results, in our case-control study of the efficacy of screening sigmoidoscopy we use the truncated product method with \( \tilde{\alpha} = 0.20 \).
Table 6: Simulated power of a sensitivity analysis of combined evidence in a case control study, where there is no unmeasured confounder and \( Pr(Z_l = 1) = 1/3 \). The response is simulated from \( \beta + t_3/\sqrt{3} \) if \( Z_l = 1 \) and \( t_3/\sqrt{3} \) if \( Z_l = 0 \). There are 1,000 narrow cases and 1,000 marginal cases with 2,000 controls. Based on 10,000 iterations. Fisher = Fisher’s combination method, tP = truncated product method with \( \tilde{\alpha} = 0.20 \).

<table>
<thead>
<tr>
<th>( \Gamma_{nm} )</th>
<th>( \Gamma_{bc} )</th>
<th>( \beta = 0 )</th>
<th>( \beta = 0.2 )</th>
<th>( \beta = 0.4 )</th>
<th>( \beta = 0.6 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fisher</td>
<td>tP</td>
<td>Fisher</td>
<td>tP</td>
<td>Fisher</td>
<td>tP</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0.0522</td>
<td>0.0525</td>
<td>0.9994</td>
<td>0.9996</td>
</tr>
<tr>
<td>1.5</td>
<td>0.0105</td>
<td>0.0152</td>
<td>0.0049</td>
<td>0.0051</td>
<td>0.9995</td>
</tr>
<tr>
<td>2</td>
<td>0.0105</td>
<td>0.0152</td>
<td>0.0004</td>
<td>0.0008</td>
<td>0.1475</td>
</tr>
<tr>
<td>2.5</td>
<td>0.0105</td>
<td>0.0152</td>
<td>0.0004</td>
<td>0.0008</td>
<td>0.0002</td>
</tr>
<tr>
<td>1.25</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.4687</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2.75</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3.5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1.5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.0017</td>
</tr>
<tr>
<td>2.5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3.5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1.75</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3.25</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2.5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3.5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2.25</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2.5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

4.6. Evidence factors with differential effect of unmeasured confounders on the factors

The individual factors in an evidence factors analysis, if biased, are hoped to be biased by different mechanisms so that, a critic would need to consider both sources of bias to explain the observed statistical significance. As discussed in Section 4.2.1, in the sigmoidoscopy study, the bias in comparing all colorectal cancer cases to controls could be due to imbalance between the two groups in healthy lifestyle of the patients, family history and also potentially due to diet. The comparison of distal cancer cases to proximal cancer cases may be biased by diet, e.g. Mediterranean diet. Hence, the main source of unmeasured confounding in the second analysis can, to some extent, also be a source of bias in the first analysis. The
following discussion delineates the logic of evidence factors analysis for such a scenario in which the sources of bias overlap for the two evidence factors but are different in their relative size between the two evidence factors.

Recall, Section 4.4.2 provides the amplification of the sensitivity parameters $\Gamma_{bc}$ and $\Gamma_{nm}$ in terms of the $\lambda$, $\delta_{bc}$ and $\delta_{nm}$. There, $u_1$ and $u_2$ are assumed to be two separate unmeasured confounds. The relation of the unmeasured confounding, $u_1$ and $u_2$, and the exposure to treatment is model by bias level $\lambda$. The relation of $u_1$ and the broad case status is modeled by the bias level $\delta_{bc}$. Finally, the relation of $u_2$ and the broad case status is modeled by the bias level $\delta_{nm}$. In the following we allow for $u_1$ and $u_2$ to be influenced by overlapping factors.

For individual $l$, let $v_{1l}$ and $v_{2l}$ be unmeasured numbers summarizing two sets of unmeasured variables so that $0 \leq v_{1l}, v_{2l} \leq 1$. We allow for both variables to bias each analysis but to have varying importance in their relationship with the outcomes. We formalize this as follows. Let $u_{1l} = \psi_1 v_{1l} + \psi_2 v_{2l}$ where $\psi_1, \psi_2 \geq 0$, $\psi_1 + \psi_2 = 1$ and $\psi_1$ is larger than $\psi_2$. Also, let $u_{2l} = \tilde{\psi}_1 v_{1l} + \tilde{\psi}_2 v_{2l}$ where $\tilde{\psi}_1, \tilde{\psi}_2 \geq 0$, $\tilde{\psi}_1 + \tilde{\psi}_2 = 1$ and $\tilde{\psi}_2$ is larger than $\tilde{\psi}_1$. The fractions $\psi_1, \psi_2, \tilde{\psi}_1$ and $\tilde{\psi}_2$ are fixed numbers. The unmeasured confounders $v_{1l}$ and $v_{2l}$ relates to the broad case status and the narrow case status by models (4.9) and (4.10) via the variables $u_{1l}$ and $u_{2l}$.

As for the relation between the unmeasured confounders $v_{1l}$, $v_{2l}$ and the observed exposure to treatment, for two units $i_1$ and $i_2$ with the same observed covariates we write, for $z_{i1} + z_{i2} = 1$

$$\Pr(Z_{i1} = z_{i1}, Z_{i2} = z_{i2} \mid C, x_{i1} = x_{i2}, Z_{i1} + Z_{i2} = 1) = \frac{\exp\{\lambda(z_{i1}\omega_1 + z_{i2}\omega_2)\}}{\exp(\lambda\omega_1) + \exp(\lambda\omega_2)}, \quad (4.13)$$

where $\omega_l = \zeta_1 v_{1l} + \zeta_2 v_{2l}$, for $l = 1, \ldots, L; \zeta_1, \zeta_2 \geq 0, \zeta_1 + \zeta_2 = 1. \quad (4.14)$

Now consider the amplification of the sensitivity parameters $\Gamma_{bc}$ and $\Gamma_{nm}$ under the model
specified by equations (4.13), (4.14) and (4.9), (4.10) with $u_{1l} = \psi_1 v_{1l} + \psi_2 v_{2l}$ and $u_{2l} = \tilde{\psi}_1 v_{1l} + \tilde{\psi}_2 v_{2l}$. This can be communicated under three different scenarios depending on the source of bias under doubt – either bias from one of $v_1$ or $v_2$, or bias from both $v_1$ and $v_2$. Assume a value of $\lambda$ in model (4.13)–(4.14). We find the parameters $\delta_{bc}$ and $\delta_{nm}$ from $\lambda$ and $\Gamma_{bc}$, $\Gamma_{nm}$. Let $\Lambda = \exp(\lambda)$, $\Delta_{bc} = \exp(\delta_{bc})$ and $\Delta_{nm} = \exp(\delta_{nm})$. Then, (i) if only $v_1$ is the bias in question, i.e. we put the restriction $v_{2,l} = v_{2,l'}$, then $\Delta_{bc} = \{(\Lambda \Gamma_{bc} - 1)/(\Lambda - \Gamma_{bc})\}^{1/\psi_1}$ and $\Delta_{nm} = \{(\Lambda \Gamma_{nm} - 1)/(\Lambda - \Gamma_{nm})\}^{1/\tilde{\psi}_1}$. This correspondence holds with $|v_{1,i1} - v_{1,i2}| = 1$.

(ii) If only $v_2$ is the bias in question, i.e. we put the restriction $v_{1,l} = v_{1,l'}$, then $\Delta_{bc} = \{(\Lambda \Gamma_{bc} - 1)/(\Lambda - \Gamma_{bc})\}^{1/\psi_2}$, $\Delta_{nm} = \{(\Lambda \Gamma_{nm} - 1)/(\Lambda - \Gamma_{nm})\}^{1/\tilde{\psi}_2}$ and $|v_{2,i1} - v_{2,i2}| = 1$. (iii) Finally, if both the confounders $v_1$ and $v_2$ are in question, then $\Delta_{bc} = (\Lambda \Gamma_{bc} - 1)/(\Lambda - \Gamma_{bc})$ and $\Delta_{bc} = (\Lambda \Gamma_{nm} - 1)/(\Lambda - \Gamma_{nm})$. This correspondence holds with $|v_{1,i1} - v_{1,i2}| = 1$ and $|v_{2,i1} - v_{2,i2}| = 1$. A closer look at these formulas immediately shows that bias parameters $\delta_{bc} = \log(\Delta_{bc})$ and $\delta_{nm} = \log(\Delta_{nm})$ changes wildly across the scenarios.

Guided by the above calculations, Figure 9 provides an illustration of the influence of unmeasured confounders on the broad case status, $\delta_{bc}$, and on the narrow case status to a marginal case status, $\delta_{nm}$. In this illustration we assume $\psi_1 = 3/4$, so that in determining a broad case status the magnitude of unmeasured confounding from $v_1$ over $v_2$ has the ratio 3:1. Whereas in determining a narrow case status to a marginal case status the magnitude of unmeasured confounding from $v_1$ over $v_2$ has the ratio 1:4, i.e., $\tilde{\psi}_1 = 1/5$. The plot considers three critics, showed in three colors, with different positions on their beliefs in the source of bias from unmeasured confounding. The first critic assumes bias only from $v_1$, the second critic assumes bias only from $v_2$ and finally the third critic assumes biases from both $v_1$ and $v_2$. The x-axis on the plot (in red) shows the amount of bias the first critic would have to assume, the y-axis on the plot (in blue) shows the amount of bias the second critic would have to assume and finally the green curves show the amount of bias the third critic would have to assume. For example, the plot highlights the situation where the critics want to explain the sensitivity of the comparisons at level $\Gamma_{bc} = 2$ and $\Gamma_{nm} = 2$, and all of them speculate $\Lambda = 4$. The first critic would have to assume biases at the amount
Figure 9: Level of bias from unmeasured confounding plotted under three speculations – bias only from $v_1$, plotted on the x-axis and in ‘red’; bias only from $v_2$, plotted on the y-axis and in ‘blue’; and biases from both $v_1$ and $v_2$, plotted in ‘green’ contours. The contours are of the function $f(\delta_{v_1}, \delta_{v_2}) = (1/\delta_{v_1} + 1/\delta_{v_2})^{-1}$. Here, $\psi_1 = 3/4$, $\psi_2 = 1/4$, $\tilde{\psi}_1 = 1/5$ and $\tilde{\psi}_2 = 4/5$. The bias levels $\delta_{bc}$ and $\delta_{nm}$ change with the speculation and the required bias level is minimized when biases from both $v_1$ and $v_2$ are assumed.

Thus, when the factors overlap but do not completely overlap in their sources of bias, evidence factors will be useful in narrowing the range of explanations for how an observed association could not be causal.
4.7. Results: efficacy of screening sigmoidoscopy

In our study of mortality from colorectal cancer and screening sigmoidoscopy, the two evidence factors analyses are summarized in Table 7. The count for screening sigmoidoscopy represent the number of individuals who had a screening procedure in 10 years before the reference date. The raw odds ratio, without controlling for any covariates, of screening sigmoidoscopy between proximal and distal cancer cases is 0.63 and that between all colorectal cancer cases and controls is 0.64. To control for important covariates we utilize the matched sets we constructed in Section 4.2.1. Using this matched sets design the p-value for efficacy of screening sigmoidoscopy for the distal colorectal cancer cases versus the proximal colorectal cancer cases is $2.3 \times 10^{-5}$, with the corresponding odds ratio 0.60 (C.I. 0.46 to 0.76)). The p-value for all cases (distal and proximal) versus the matched controls is $5.0 \times 10^{-11}$, with odds ratio 0.62 (C.I. 0.54 to 0.72)) (this result is similar to previously reported odds ratios ([Atkin et al., 2010; Segnan et al., 2011]).

We further conduct a sensitivity analysis to assess whether possible covariates which were not controlled for in our study may have been the reason behind the observed association above. Being consistent with the notation of Section 4.4 we consider two sensitivity parameters $\Gamma_{nm}$ and $\Gamma_{bc}$ for the two comparisons. A value of 1 for a sensitivity parameter would say that there is no bias from unmeasured confounding in the respective comparison, and the higher the value is of the parameter, the bigger is the bias. Figure 10 shows the bias levels where the combined evidence for a beneficial effect of screening sigmoidoscopy is sensitive. The p-value upper bounds for each bias level of the two evidence factors are combined using the truncated product method with $\tilde{\alpha} = 0.20$. As can be seen in this plot,

<table>
<thead>
<tr>
<th>Table 7: Screening sigmoidoscopy and colorectal cancer summary data.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td>Distal cancer cases</td>
</tr>
<tr>
<td>---------------------</td>
</tr>
<tr>
<td>No screening sigmoidoscopy</td>
</tr>
<tr>
<td>Screening sigmoidoscopy</td>
</tr>
<tr>
<td>Odds ratio from matched sets</td>
</tr>
<tr>
<td>p-value from matched sets</td>
</tr>
</tbody>
</table>
Figure 10: Sensitivity analysis of the efficacy of screening sigmoidoscopy in reducing mortality from colorectal cancer. The darker gray color represents the bias levels where the combined evidence for a beneficial effect of screening sigmoidoscopy is sensitive.

only a substantial amount of bias in both comparisons could explain the observed association in the data, if in fact the null hypothesis is true. For example, with a maximum bias of $\Gamma_{nm} = 1.4$ in the comparison of distal cancer cases to proximal cancer cases, the combined evidence is sensitive only when the bias in the second comparison of all colorectal cases to the controls is larger than $\Gamma_{bc} = 1.45$. The overall evidence remains insensitive for $\Gamma_{nm} = 2$ when $\Gamma_{bc} \leq 1.35$. Thus the overall evidence for the efficacy of the procedure is strengthened compared to evidence from an analysis that only looks at the screening rates between all colorectal cancer cases and controls.
4.8. Discussion

In this chapter we have developed evidence factors in a case-control study in which there is a narrow and a broad case definition. These evidence factors are formed by two sets of comparisons, the first one comparing narrow cases to marginal cases and the second one comparing all cases to controls. Use of these evidence factors in a case-control study can provide better insight into the study especially in a discussion and analysis of possible bias in the study.

In the sigmoidoscopy study, we pair matched narrow cases to marginal cases and included their controls in the matched sets, and then put the remaining marginal cases in matched sets with their controls. Other matching methods could be used, e.g., full matching (Hansen, 2004) or variable ratio matching (Ming and Rosenbaum, 2000; Pimentel et al., 2015).

4.8.1. Other examples with multiple case definitions

In certain diseases, like cancer in the body of the uterus, atherosclerosis, hypertension and mental illness, multiple case definitions are considered or often necessary (Acheson, 1979; Cole, 1979; Cohen et al., 2005). Some other specific studies where multiple case definitions have been considered are discussed here. These studies illustrate various ways to design a broad case vs. narrow case distinction in case-control studies. In a study to assess whether statin causes peripheral neuropathy Gaist et al. (2002) classify the neuropathy cases as definite and nondefinite cases of idiopathic peripheral neuropathy based on the intensity of the symptom and the quality of the clinical information. In the terminology of the present chapter the definite cases would be the narrow cases where the association, if present, would be stronger compared to the marginal cases, i.e., the nondefinite cases. Small et al. (2013) use an illustrative case-control study for physical abuse by parents in childhood and tendency for more anger in adulthood. In this study the cases were split in two definitions based on whether or not anger score was on a higher range. Here, a case on a higher quantile of anger score could be defined as a narrow case. As a final
example, in an effort to understand association between genetic traits and cerebral malaria, Small et al. (2017) consider cerebral malaria cases with and without retinopathy. The World Health Organization (WHO) defines a child as having cerebral malaria when the child is in a coma (cannot localize a painful stimulus), has malaria parasites in his or her blood and has no other known cause of the coma. This definition is not specific as hospitals in malaria-endemic areas often lack diagnostic facilities to identify non-malarial causes of coma and many children in malaria endemic areas have non-symptomatic malaria infections. There are characteristic retinal abnormalities (retinopathy) that increase the specificity of a cerebral malaria diagnosis (Taylor et al., 2004). Cerebral malaria cases with such retinal abnormalities could be considered as narrow cases and those without the retinal abnormalities could be considered as marginal cases.
CHAPTER 5: Reinforced designs for observational studies of treatment effects:
Multiple instruments plus control groups as evidence factors

5.1. Introduction: active, effective concern for bias in observational studies

5.1.1. The problem and passive, ineffective solutions
Absent random assignment, an association between treatment received and outcome exhibited may not reflect an effect caused by the treatment, but rather some bias in the way people were selected for treatment. Biases visible in observed pretreatment covariates are removed by adjustments, such as stratification combined with covariance adjustment, but there is always concern that people who look similar in terms of observed covariates may differ in covariates not observed. In published empirical work on the effects of treatments on human subjects, the most common way to address this issue is also the most passive and ineffective: the possibility that unmeasured bias invalidates the study’s conclusion is mentioned in the discussion section (Rosenbaum 1991).

One active, enlightening approach uses the available data to quantify the extent of uncertainty due to unmeasured covariates. A sensitivity analysis asks about the magnitude of unmeasured bias in treatment assignment that would need to be present to materially alter the causal conclusions. This approach can be effective: it distinguished the evidence that smoking causes lung cancer from the presumably spurious association between smoking and cirrhosis of the liver — many smokers drink to excess. The strong association between heavy smoking and lung cancer can only be explained away by enormous biases in who chooses to smoke. See Cornfield et al. (1959) for the first sensitivity analysis in an observational study, Hammond (1964, Table 10) for the many associations between smoking and diseases, including lung cancer and cirrhosis, and Rosenbaum (2002, §4.3.2) for reanalysis of Hammond’s data. For various perspectives on sensitivity analysis in observational studies, see Hosman et al. (2010), Gilbert et al. (2003), Rudolph and Stuart (2017), Schwartz et al. (2012) and Yu and Gastwirth (2005).
An active but ineffective approach performs several closely parallel analyses of the same data, several analyses that all assume the absence of unmeasured bias in treatment assignment, using the term “sensitivity analysis” to designate to these several analyses. If several similar analyses of the same data all make the same assumption that unmeasured biases are absent, and if these analyses concur, then that is evidence that the investigator did lots of analysis before selecting results for publication, not evidence that the association between treatment and outcome is an effect caused by the treatment. Rubin (2007) was sharply critical of investigators who immediately dive in to analyzing outcomes, putting little or no effort into design before examining outcomes; specifically, he questioned the objectivity of such investigators when selecting a few of many parallel analyses for publication.

An instrument or instrumental variable is a bit of randomized or haphazard encouragement to accept a treatment in a context in which treatment assignment is often deliberate or thoughtful, hence potentially biased; see Angrist, Imbens and Rubin (1996). In their example, the Vietnam War draft lottery was randomized, and it pushed some men into military service who would not otherwise have served, but some men volunteered while others dodged the draft, so the lottery was a random push amid other, potentially biased, considerations that decided military service. The typical instrument is less compelling than the draft lottery because it is not actually randomized, and it makes assumptions that are no less speculative than the assumption that unmeasured biases are absent. As a single treatment might be encouraged by various instruments, it is not a foregone conclusion that analyses with different instruments will concur; so, it is informative if they do concur (Imbens and Rosenbaum 2005, §1.1). For discussion of instruments from several perspectives, see Angrist et al. (1996), Baiocchi, Cheng and Small (2014), Brookhart et al. (2010), Hogan and Lancaster (2004), Kang (2016), Keele and Morgan (2016), Larcker and Rusticus (2010), Li et al. (2015), Lu and Marcus (2012) and Small (2007). For an example, see Neuman et al. (2014).

Multiple analyses provide evidence about unmeasured biases if: (i) certain biases that would
invalidate one analysis do not bias another analysis, (ii) each analysis is insensitive to small or moderate biases of the type that might invalidate that analysis, and (iii) these several analyses would be nearly statistically independent if the treatment had no effect. Analyses of this type are called evidence factors (Rosenbaum 2010, 2011, 2017). Because these analyses are affected by different types of unmeasured biases and are nearly independent despite using the same data, it is far from a forgone conclusion that the analyses will concur. See Zubizarreta et al. (2012) for an example in which two evidence factors do not concur, thereby providing evidence that at least some associations are spurious, not causal. See Zhang et al. (2011) for an example in which two evidence factors concur.

5.1.2. Goal: Several instruments and direct comparisons as evidence factors

Our goal is to design observational studies to use several instruments plus direct comparisons of treated and control groups as evidence factors. Typically, when several instruments are available, investigators employ them in a joint analysis, such as two-stage least squares, so if any of the instruments fails to satisfy the many assumptions required of an instrument, then the joint analysis is compromised. However, see Kang et al. (2016) for an approach that tolerates some invalid instruments. In contrast, we use several instruments one at a time in such a way that failure of the assumptions for one instrument does not, by itself, invalidate analyses with other instruments.

A final evidence factor directly compares treated and control groups. Some scientific fields presume that direct comparisons are more easily dismissed as bias than are comparisons using instruments, but this presumption is not true in general. As we show, in a single sampling situation, a direct comparison may be insensitive to large biases, while the analysis with an instrument may be sensitive to small biases, so the direct comparison may provide the most compelling evidence, as it did in the case of smoking and lung cancer. We demonstrate this theoretically by calculating design sensitivities.
5.1.3. **Outline**

To fix ideas, §5.2 briefly introduces the application, to which we return in §5.6. Section 5.3 defines notation and recalls a few basic ideas about causal inference and sensitivity analysis. Key results are in §5.4: these demonstrate that several subanalyses are, indeed, evidence factors. The proposed method is evaluated in §5.5 in terms of design sensitivity, with some surprises about the relative safety of instruments and direct comparisons. The chapter concludes with a brief summary and discussion in §5.7.

5.2. How effective are Catholic schools compared to public school?

Are private Catholic high schools more effective than public high schools? This perennial question is revisited in current debates about charter schools and public subsidies for private education. It is not, however, an easy question to answer. Paying more to attend Catholic school may signify a parent’s concern or commitment to education, which may affect outcomes in many ways. Even after adjustment for educational and socioeconomic covariates, a direct comparison of students in public and Catholic school may, therefore, be biased. The empirical literature contains: (i) attempts to use the geographic accessibility of Catholic schools as an instrument for attending Catholic schools, (ii) attempts to use “being a Catholic” as an instrument, (iii) direct comparisons of students in Catholic and public high schools, (iv) sharp conflict about which, if any, of these approaches yields valid inferences about the effects caused by Catholic schools. See Altonji, et al. (2005), Coleman (1982), Goldberger and Cain (1982), Kim (2011), and Neal (1997) for several perspectives and analyses.

Rather than select one analysis and assert that it is valid, we develop three evidence factors, three (nearly) statistically independent analyses of the same data, each dependent upon very different assumptions for its validity. Because these analyses are independent, they do not repeat one another, and concurrence among the analyses is far from a foregone conclusion. Because certain biases that would invalidate one analysis have no effect on another analysis,
concurrency would weaken some claims that biases produced the ostensible treatment effects. To the extent that each analysis, each factor, is insensitive to the type of bias that could invalidate that factor, there is further weakening of claims that bias accounts for ostensible effects. Conversely, the absence of concurrence and sensitivity to small biases are warnings that bias could readily explain ostensible effects.

Using data from the Wisconsin Longitudinal Study, we will examine income from wages and salary in 1974 for 4450 male students who completed high school in Wisconsin in 1957. Table 8 depicts the structure of the three factors, essentially (i)-(iii) above. The 4450 students divide into 1501 students from urban Wisconsin and 2949 from rural Wisconsin, and presumably because Catholic schools are more accessible in urban areas, 22% of urban students attended Catholic school, while only 6% of rural students attended Catholic school. So the first analysis uses urban/rural as an instrument for Catholic education. The second analysis compares children in urban areas to other children in urban areas, and children in rural areas to other children in rural areas, so the second analysis views urban/rural as a covariate, not an instrument. In urban areas, roughly half the students were Catholic, and 44% of Catholics attended Catholic high schools, while none of the non-Catholics attended Catholic high schools. In rural areas, Catholic students were in the minority, and 17% of Catholics attended Catholic high schools, while only 2 of the non-Catholics, or less than half a percent, attended Catholic high schools. The second analysis views Catholic religion as an instrument, and urban/rural as a covariate. The third analysis has no instrument: it directly compares students attending Catholic and public schools, viewing both urban/rural and Catholic/non-Catholic as covariates.

The example is discussed in detail in §5.6.
Table 8: Counts and percents attending Catholic school for two potential instruments and a direct comparison of Catholic and public schools.

<table>
<thead>
<tr>
<th>Group</th>
<th>Count</th>
<th>% Attending Catholic School</th>
</tr>
</thead>
<tbody>
<tr>
<td>Urban Catholic</td>
<td>1501</td>
<td>22</td>
</tr>
<tr>
<td>Urban Public</td>
<td>414</td>
<td>0</td>
</tr>
<tr>
<td>Other Catholic</td>
<td>760</td>
<td>0</td>
</tr>
<tr>
<td>Other Public</td>
<td>760</td>
<td>0</td>
</tr>
<tr>
<td>Rural Catholic</td>
<td>2949</td>
<td>6</td>
</tr>
<tr>
<td>Rural Public</td>
<td>868</td>
<td>0</td>
</tr>
<tr>
<td>Other Catholic</td>
<td>1904</td>
<td>2</td>
</tr>
<tr>
<td>Other Public</td>
<td>1902</td>
<td>0</td>
</tr>
<tr>
<td>Total</td>
<td>4450</td>
<td>5.3.</td>
</tr>
</tbody>
</table>

5.3. Notation and background: several instruments plus a direct comparison

5.3.1. Notation: strata, covariates, outcomes, treatment and instruments
There are $I$ strata, $i = 1, \ldots, I$, with $n_i$ individuals $ij$ in stratum $i$, $j = 1, \ldots, n_i$, and $N = \sum_{i=1}^{I} n_i$ individuals in total. There are $K$ binary, 1 or 0, indicators, $Z_{ijk}$, $k = 1, \ldots, K$. The first $K - 1$ indicators are possible instruments for indicator $K$ which is the active treatment. In §5.2: (i) $Z_{ij1} = 1$ for urban residence, $Z_{ij1} = 0$ for rural residence; (ii) $Z_{ij2}$ distinguishes Catholics, $Z_{ij2} = 1$, from others, $Z_{ij2} = 0$; (iii) $Z_{ij3}$ distinguishes attending a Catholic high school, $Z_{ij3} = 1$, from attending a public school, $Z_{ij3} = 0$.

Individual $ij$ has an observed covariate $x_{ij}$ controlled by stratification, so $x_{ij} = x_{ij'}$ for $1 \leq j < j' \leq n_i$. There is concern about an unobserved covariate $u_{ijk}$, $k = 1, \ldots, K$, not controlled by stratifying on $x_{ij}$. The notation permits a different unobserved $u_{ijk}$ for each $Z_{ijk}$, but there is no requirement that they be distinct; that is, the situation with $u_{ij1} = \cdots = u_{ijK} = u_{ij}$, say, is simply a special case.

If the exclusion restriction of Angrist et al. (1996) held for all of the first $K - 1$ indicators, then individual $ij$ would exhibit response $r_{Tij}$ if $Z_{ijk} = 1$ or response $r_{Cij}$ if $Z_{ijk} = 0$. In fact, our $K$ analyses do not assume that the exclusion restriction holds for all $K - 1$ potential instruments, but rather assume much less. The analysis that uses $Z_{ijk}$ as if it were an instrument assumes that the exclusion restriction holds for $Z_{ijk}, \ldots, Z_{ijk-1}$ when we
compare individuals who are the same in terms of $Z_{ij1}, \ldots, Z_{ijk-1}$, so the exclusion restriction may not hold for $Z_{ij1}, \ldots, Z_{ijk-1}$. The direct comparison of treated and control individuals, $Z_{ijK} = 1$ versus $Z_{ijK} = 0$, does not assume any exclusion restriction, simply adjusting for $Z_{ij1}, \ldots, Z_{ijk-1}$.

There are $K$ partial assignment vectors, $A_{ijk} = (Z_{ij1}, \ldots, Z_{ijk})$ for $k = 1, \ldots, K$, and a matrix $A_k$ whose $N$ rows are the $A_{ijk}$, so $A_k$ records assignments up to step $k$ for all $N$ individuals. It is notationally convenient to define $A_{ij0} = \emptyset$, so that $A_{ij,k-1}$ is well defined for $k = 1$, but conditioning on $A_{ij0}$ means that no part of $A_{ij} = (Z_{ij1}, \ldots, Z_{ijk})$ is actually being conditioned upon. Let $A_k$ be the set containing the $2^k$ vectors of dimension $k$ with 1 or 0 coordinates. The vector, $A_{ijk} = (Z_{ij1}, \ldots, Z_{ijk})$, can take on $2^k$ possible values $a \in A_k$. The entire study amalgamates $K$ partial studies, where study $k$ fixes the $2^{k-1}$ values $a \in A_{k-1}$ of $A_{ij,k-1}$, studies the effects of variations in $Z_{ijk}$, and lets $(Z_{ij,k+1}, \ldots, Z_{ijk})$ fluctuate as it will. In Table 8, there are $K = 3$ partial studies. At assignment $k$, individual $ij$ has $2^k$ potential outcomes, $r_{ija}$ with $a \in A_k$, so each step is an instance of the Neyman (1923) - Rubin (1974) notation for causal effects. However, at step $k$, we are interested in comparing $r_{ija}$ and $r_{ija'}$ for each pair $(a, a')$ with $a, a' \in A_k$ such that $a = (a_1, \ldots, a_{k-1}, 1)$ and $a' = (a_1, \ldots, a_{k-1}, 0)$, so $a$ and $a'$ differ only in the last, $k$th, coordinate; moreover, the $k$th partial study is focused on this comparison. At assignment $k$, Fisher’s hypothesis $H_k$ of no effect of assignment $k$ asserts that $r_{ija} = r_{ija'}$ for each pair $(a, a')$ with $a, a' \in A_k$ such that $a = (a_1, \ldots, a_{k-1}, 1)$ and $a' = (a_1, \ldots, a_{k-1}, 0)$. In total, individual $ij$ has $\sum_{k=1}^K 2^k = 2^{K+1} - 2$ potential outcomes $r_{ija}$ for $A_{ijk} = a \in A_k$, for $k = 1, \ldots, K$, which we collect in a vector $r_{ij}$ of dimension $2^{K+1} - 2$. Later, in §5.3.3, when we impose a “partial exclusion condition” on $r_{ij}$, the potential complexity of $r_{ij}$ will be greatly reduced. In effect, the partial exclusion restriction will say that for people who are the same in terms of $(Z_{ij1}, \ldots, Z_{ijk-1}),$ the vector $(Z_{ijk}, \ldots, Z_{ijk,K-1})$ affects $r_{ij}$ only indirectly by altering $Z_{ijkK}$; see, again, Table 8. Ultimately, we observe only one coordinate of $r_{ij}$, namely $R_{ij} = r_{ija}$ for $A_{ijK} = a$ for $a \in A_K$, so much of $r_{ij}$ is inaccessible to us. Write $R = (R_{i1}, \ldots, R_{I,nj})^T$ for the $N$-dimensional vector of observed responses.
Concerning notation, note that a vector $\mathbf{a} \in \mathcal{A}_k$ has dimension $k$, so the notation $r_{ija}$, $\mathbf{a} \in \mathcal{A}_k$, is well defined without mentioning $k$.

Write $\mathcal{F} = \{r_{ij}, x_{ij}, u_{ijk}, i = 1, \ldots, I, j = 1, \ldots, n_i, k = 1, \ldots, K\}$. Conditionally given $\mathcal{F}$, distinct individuals, say $ij$ and $i'j'$, are assumed to have independent values of the $K$-dimensional assignment vector, $\mathbf{A}_{ij} = (Z_{ij1}, \ldots, Z_{ijK})$ and $\mathbf{A}_{i'j'}$. Write the $N$-dimensional vector of assignments at step $k$ as $\mathbf{Z}_k = (Z_{11k}, \ldots, Z_{I,n_1,k})^T$, for $k = 1, \ldots, K$.

### 5.3.2. Treatment assignment in $K$ steps

Consider the model for treatment assignment

$$
\Pr (Z_{ijk} = 1 \mid \mathcal{F}, \mathbf{A}_{ij,k-1}) = \frac{\exp \{\kappa_k (x_{ij}, \mathbf{A}_{ij,k-1}) + \gamma_k u_{ijk}\}}{1 + \exp \{\kappa_k (x_{ij}, \mathbf{A}_{ij,k-1}) + \gamma_k u_{ijk}\}},
$$

for all $i, j, k$, with $0 \leq u_{ijk} \leq 1$,

where $\kappa_k (\cdot)$ is an unknown function and $\gamma_k \geq 0$ is an unknown sensitivity parameter. This model says that $Z_{ijk}$ depends in an entirely arbitrary way on the observable $(x_{ij}, \mathbf{A}_{ij,k-1})$, but otherwise depends upon $\mathcal{F}$ only through $u_{ijk}$. Model (5.1) says that two individuals, $j$ and $j'$, with the same $(x_{ij}, \mathbf{A}_{ij,k-1})$ differ in their conditional odds of treatment at step $k$ by at most $\Gamma_k = \exp (\gamma_k) \geq 1$, namely

$$
\frac{1}{\Gamma_k} \leq \frac{\Pr (Z_{ijk} = 1 \mid \mathcal{F}, \mathbf{A}_{ij,k-1}) \Pr (Z_{ij'k} = 0 \mid \mathcal{F}, \mathbf{A}_{ij',k-1})}{\Pr (Z_{ij'k} = 1 \mid \mathcal{F}, \mathbf{A}_{ij',k-1}) \Pr (Z_{ijk} = 0 \mid \mathcal{F}, \mathbf{A}_{ij,k-1})} \leq \Gamma_k
$$

whenever $(x_{ij}, \mathbf{A}_{ij,k-1}) = (x_{ij'}, \mathbf{A}_{ij',k-1})$.

If $\gamma_k = 0$, then there is no bias in assignment at step $k$, in the sense that everyone with the same observed $(x_{ij}, \mathbf{A}_{ij,k-1})$ has the same probability of $Z_{ijk} = 1$, so assignment is ignorable at step $k$. Write $\mathbf{u}_k = (u_{11k}, \ldots, u_{I,n_1,k})^T$ and $\mathcal{U} = [0, 1]^N$ for the $N$-dimensional unit cube.

If $\mathcal{S}$ is a finite set, write $|\mathcal{S}|$ for the number of elements of $\mathcal{S}$.

If $\mathbf{a} \in \mathcal{A}_{k-1}$ is a $(k - 1)$-dimensional vector of 0s and 1s, let $\mathcal{T}_{i\mathbf{a}} \subseteq \{1, \ldots, n_i\}$ be the subset
of individuals in stratum \( i \) with \( A_{ijk-1} = a \), let \( n_i,a = |T_{i,a}| \) and \( m_{i,a} = \sum_{j \in T_{i,a}} Z_{ijk} \). Write \( m_{ik} \) for the vector of dimension \( 2^{k-1} \) whose coordinates are the \( m_{i,a} \) for the \( 2^{k-1} \) possible values of \( a \in A_{k-1} \). Again, it is convenient to use the same notation for \( k = 1 \), where \( k = 1 \), \( a = \emptyset \), \( T_{i,a} = \{1, \ldots, n_i\} \), \( n_i,a = n_i, m_{i,a} = \sum_{j = 1}^{n_i} Z_{ij1} \). For \( a \in A_{k-1} \), let \( Z_{i,a,\ell} \) be the value of \( Z_{ijk} \) for individual \( \ell \in T_{i,a} \), let \( Z_{i,a} \) be the corresponding column vector of dimension \( n_i,a \) with \( m_{i,a} \) ones and \( n_i,a - m_{i,a} \) zeros, and let \( u_{i,a} \) be the column vector of dimension \( n_i,a \) with \( m_{i,a} \) ones. As there is typically uncertainty and contention about whether the exclusion restriction actually holds for possible instruments, we introduce a partial exclusion restriction. In effect, this condition says that some of the \( Z_{ijk} \) satisfy an exclusion restriction, others do not, consistent with some \( Z_{ijk} \) being instruments, while other \( Z_{ijk} \) require adjustments, similar to the adjustments for covariates.

\[ \Pr (Z_{i,a} = z \mid F, A_{k-1}, m_{ik}) = \frac{\exp(\gamma_k z^T u_{i,a})}{\sum_{b \in Z_{i,a}} \exp(\gamma_k b^T u_{i,a}) } \text{ for } z \in Z_{i,a}, u_k \in U, \quad (5.2) \]

because \( \kappa_k (x_{ij}, A_{ij,k-1}) \), though unknown, takes the same value for all individuals \( ij \) with \( j \in T_{i,a} \). Moreover, the \( 2^{k-1} \) distinct \( Z_{i,a} \) for the \( 2^{k-1} \) values of \( a \) are conditionally independent of each other given \( F, A_{k-1}, m_{ik} \). For fixed \( k \), as \( a \in A_{k-1} \) varies over its \( 2^{k-1} \) possible values, model (5.2) is a conventional model for stratified, treatment/control sensitivity analyses with \( I \times 2^{k-1} \) strata; see Rosenbaum and Krieger (1990), Rosenbaum and Small (2017) and Rosenbaum (2018). If \( \gamma_k = 0 \), then (5.2) becomes random assignment, \( \Pr (Z_{i,a} = z \mid F, A_{k-1}, m_{ik}) = |Z_{i,a}|^{-1} \) for each \( z \in Z_{i,a} \), so model (5.2) permits one of the \( K \) steps to be free of bias from \( u_{ijk} \) while the other \( K - 1 \) steps are biased.

5.3.3. The partial exclusion restriction

As there is typically uncertainty and contention about whether the exclusion restriction actually holds for possible instruments, we introduce a partial exclusion restriction. In effect, this condition says that some of the \( Z_{ijk} \) satisfy an exclusion restriction, others do not, consistent with some \( Z_{ijk} \) being instruments, while other \( Z_{ijk} \) require adjustments, similar to the adjustments for covariates.
If \((Z_{ij1}, \ldots, Z_{ijK})\) were \(K\) two-level treatments in a \(2^K\) factorial experiment, then each individual would have \(2^K\) potential outcomes depending upon the \(2^K\) ways that the \(K\) treatments \((Z_{ij1}, \ldots, Z_{ijK})\) might be set. In contrast, the assumption that \((Z_{ij1}, \ldots, Z_{ijK-1})\) are \(K-1\) valid instruments for an active treatment \(Z_{ijK}\) entails, among other things, an exclusion restriction which says there are only two potential outcomes, \(r_{Tij}\) if \(Z_{ijK} = 1\) or \(r_{Cij}\) if \(Z_{ijK} = 0\). In words, the exclusion restriction says \((Z_{ij1}, \ldots, Z_{ijK-1})\) may push an individual \(ij\) towards treatment, \(Z_{ijK} = 1\), or towards control, \(Z_{ijK} = 0\), but it is only the active treatment, \(Z_{ijK}\), that affects outcomes. In §5.2, the exclusion restriction says that being Catholic or being in an urban area affects your educational outcomes only indirectly to the extent to which it shifts you from a public to a Catholic high school, from \(Z_{ijK} = 0\) to \(Z_{ijK} = 1\). This may or may not be true. It is common to criticize conclusions based on a purported instrument by claiming that the exclusion restriction does not hold, for instance that Catholics should be compared to other Catholics because being Catholic is directly relevant to educational outcomes quite apart from attending Catholic school. To address such concerns, Definition 9 entertains the possibility that the exclusion restriction holds for parts of \((Z_{ij1}, \ldots, Z_{ijK})\) but not all of it.

**Definition 9** Let \(\mathcal{K} \subseteq \{1, 2, \ldots, K\}\), and let \(k\) be the smallest element in \(\mathcal{K}\). The partial exclusion restriction holds for \(\mathcal{K}\) if, with \(A_{ij,k-1} = (Z_{ij1}, \ldots, Z_{ij,k-1})\) fixed by conditioning, each individual \(ij\) has two potential outcomes depending upon the value of \(Z_{ijK}\), namely \(r_{Tij}\) if \(Z_{ijK} = 1\) or \(r_{Cij}\) if \(Z_{ijK} = 0\).

A partial exclusion restriction places a restriction on \(r_{ij}\), saying that \(r_{ija}\) for \(a = (a_1, \ldots, a_K)\) may vary with \((a_1, \ldots, a_{k-1})\) and \(a_K\), but not with \((a_k, \ldots, a_{K-1})\). More specifically, this restriction says: if \(a, a' \in \mathcal{A}_K\) with \(a = (a_1, \ldots, a_{k-1}, a_k, \ldots, a_{K-1}, a_K)\) and \(a' = (a_1, \ldots, a_{k-1}, a'_k, \ldots, a'_{K-1}, a_K)\), then \(r_{ija} = r_{ija'}\), and we write \(r_{ija} = r_{Cij}\) if \(a_K = 0\) or \(r_{ija} = r_{Tij}\) if \(a_K = 1\) in an analysis that fixes \(A_{ij,k-1} = (Z_{ij1}, \ldots, Z_{ij,k-1})\) by conditioning; however, \(r_{ija}\) may vary with \((a_1, \ldots, a_{k-1})\), so this notation is meaningful only with \(A_{ij,k-1} = (Z_{ij1}, \ldots, Z_{ij,k-1})\) fixed, as it would be fixed if it were a covariate rather than an
instrument. If a partial exclusion restriction holds for \( K \) and \( K' \subset K \), then a partial exclusion restriction holds for \( K' \). If the partial exclusion restriction holds for \( K \subseteq \{1, 2, \ldots, K\} \), then Fisher’s hypothesis \( H_k \) of no effect at assignment \( k \) is the same null hypothesis for each \( k \in K \), namely \( H_0 : r_{Tij} = r_{Cij}, \forall i, j \).

To clarify Definition 9, consider a few special cases. If \( K = \{1, 2, \ldots, K\} \), then partial exclusion is no different from the usual exclusion restriction for the \( K - 1 \) instruments jointly. If \( K = \{K\} \), then partial exclusion is simply the Neyman-Rubin notation for causal effects, with \( A_{ij, K-1} = (Z_{ij1}, \ldots, Z_{ij, K-1}) \) fixed as covariates rather than instruments, that is, with \( I \times 2^{K-1} \) strata defined by \( (x_{ij}, A_{ij, K-1}) \). In §5.2, partial exclusion for \( K = \{2, 3\} \) would be the usual exclusion restriction for ‘being Catholic,’ \( Z_{ij2} \), if ‘being urban or rural,’ \( Z_{ij1} \), were controlled as a covariate, that is, with \( I \times 2 \) strata. In §5.2, passing from \( K = \{1, 2, 3\} \) to \( K = \{2, 3\} \) entails two changes: first, ‘being urban or rural,’ \( Z_{ij1} \), is no longer assumed to satisfy the exclusion restriction; second, ‘being Catholic,’ \( Z_{ij2} \), is assumed to satisfy the exclusion restriction only after adjustment for ‘being urban or rural.’

Definition 9 mentions a set \( K \) but makes use only of the smallest \( k \in K \): whether the partial exclusion restriction holds for the set \( K \) depends only on its smallest element. This will be convenient later, in particular in Definitions 10 and 11. There is more to a valid instrument than the exclusion restriction; the instrument must be randomized in a certain sense; see Definition 10. An analysis might omit a potential instrument, \( Z_{ij^k} \), because of concern that \( Z_{ij^k} \) is not randomized. The analysis for \( K = \{1, 2, 3\} \) and \( K' = \{1, 3\} \) will entail the same partial exclusion restriction because the minimal element is \( k = 1 \) in both \( K \) and \( K' \), but the analysis for \( K' = \{1, 3\} \) will not use \( Z_{ij2} \), so it will not require that \( Z_{ij2} \) be randomized. If the analyses for \( K \) and \( K' \) concur, then we are less worried that a doubtful assumption about \( Z_{ij2} \) is critical to the study’s conclusions.

Suppose that the partial exclusion restriction holds for \( K \), and let \( k \) be the smallest element in \( K \). Then, by definition, an analysis that fixes \( A_{ij, k-1} = (Z_{ij1}, \ldots, Z_{ij, k-1}) \) by conditioning, by stratifying on \( A_{ij, k-1} \), has two potential outcomes, \( (r_{Tij}, r_{Cij}) \), for individual \( ij \).
depending upon the value of $Z_{ijK}$. In this case, we may entertain the null hypothesis of a shift effect, $H^\beta_k : r_{Tij} = r_{Cij} + \beta$, so that $R_{ij} - \beta Z_{ijK} = r_{Cij}$ satisfies the null hypothesis of no effect. In the conventional way, we may invert a test of no effect to obtain a $1 - \alpha$ confidence interval for $\beta$, testing every possible $\beta$ and retaining for interval the values of $\beta$ not rejected at level $\alpha$. If the partial exclusion restriction holds for $K$, then the hypothesis $H^\beta_\ell : r_{Tij} = r_{Cij} + \beta$ is the same hypothesis for every $\ell \geq k$ and in particular for every $\ell \in K$. As a consequence, we can ask whether several analyses concur in their assessment of the evidence against a specific value of $\beta$; that is, we are not restricted to asking about whether analyses concur in testing no effect.

5.3.4. Test statistics and sensitivity analyses

For each $k$, there is a stratified comparison of individuals with $Z_{ijk} = 1$ or $Z_{ijk} = 0$ within $I \times 2^{k-1}$ strata defined by the $I$ original strata based on $x_{ij}$, together with the $2^{k-1}$ strata defined by $A_{ij,k-1} = (Z_{ij1}, \ldots, Z_{ij,k-1})$. A statistic testing $H_0$, say $T_k = t_k(Z_k, R)$, at step $k$ is a function of the observed responses, $R$, and the treatment assignments, $Z_k$, at step $k$. In principle, $T_k$ may depend also on the $x_{ij}$, the $A_{ij,k-1}$, and the $m_{ik}$, but the notation does not indicate this explicitly. In the analyses in the current chapter, $T_k$ is the weighted combination of stratum-specific Wilcoxon rank sum statistics suggested by van Elteren (1960), but the Hodges-Lehmann aligned rank test or tests based on M-statistics are practical alternatives. See Lehmann (1975, §3) for comparison of several stratified tests.

Consider the $K$ steps one at a time, delaying consideration of their interdependence to §5.4. At step $k$, assume the partial exclusion restriction holds for $K \subseteq \{k, k+1, \ldots, K\}$ with $k \in K$. Then the sensitivity analysis at each step $k$ has a conventional form, and can be analyzed in a conventional way, as in Rosenbaum and Small (2017) and Rosenbaum (2018). If (5.2) were true at step $k$, and if $\gamma_k = 0$, then Fisher’s hypothesis $H_0$ of no effect would be tested by comparing $T_k = t_k(Z_k, R)$ to its stratified randomization distribution. For $\gamma_k = \log(\Gamma_k) > 0$, there is a $P$-value testing $H_0$ at the true value of $u_k$ obtained, from elementary principles, by multiplying $I \times 2^{k-1}$ expressions of the form (5.2) over the
I \times 2^{k-1} \text{ strata to obtain the probability of a single possible value } z_k \text{ of } Z_k, \text{ then summing such terms over all } z_k \text{ such that } t_k(z_k, R) \geq t_k(Z_k, R). \text{ This true } P\text{-value is not available to us because } u_k \text{ is not observed, so we find the maximum such } P\text{-value over } u_k \in U, \text{ say } P_{k, \Gamma_k}. \text{ To make the computations practical, a large sample approximation is used in place of the exact distribution in (5.2). If } \gamma_k = 0, \text{ this maximum } P\text{-value is the randomization } P\text{-value, but as } \gamma_k \to \infty \text{ the bound } P_{k, \Gamma_k} \to 1, \text{ reflecting the familiar fact that an association, no matter how strong, does not logically entail causation — sufficiently large biases can explain away an association. The practical question is quantitative, not logical: How much bias, measured by } \Gamma_k = \exp(\gamma_k), \text{ would need to be present to render } H_0 \text{ plausible?}

5.4. Evidence factors: Combining the } K \text{ steps

5.4.1. Valid or biased assignment

The validity of an instrument requires both the exclusion restriction and a type of randomness. Again, we want to entertain the possibility that some of the } K \text{ comparisons are valid and others are not.

\textbf{Definition 10} Let } K \subseteq \{1, 2, \ldots, K\}. \text{ The instrumental and direct comparisons in } K \text{ are valid if: (i) partial exclusion holds for } K, \text{ and (ii) treatment assignment is governed by (5.2) with } \gamma_k = 0 \text{ for each } k \in K.

If the instrumental and direct comparisons in } K \text{ were valid, then we could perform } |K| \text{ separate valid tests of } H_0 \text{ using stratified randomization inference, one for each } k \in K. \text{ For instance, in } \S 5.2, \text{ if the instrumental and direct comparisons in } K = \{1, 2\} \text{ were valid, then: (i) using “urban/rural”, } Z_{ij1}, \text{ alone as an instrument would yield a valid randomization test of } H_0, \text{ (ii) using “being Catholic”, } Z_{ij2}, \text{ as an instrument within } 2 \times I \text{ strata that controlled for “urban/rural”, } Z_{ij1}, \text{ would yield a valid randomization test of } H_0, \text{ but (iii) the direct comparison, } Z_{ij3}, \text{ of students in Catholic and public school, adjusting for } Z_{ij1} \text{ and } Z_{ij2}, \text{ may be biased by } u_{ij3}.

In the absence of actual randomization, we cannot be sure a comparison is valid. Definition
11 refers to a measured degree of bias in some comparisons, with the possibility that other comparisons are severely biased.

**Definition 11** Let $\mathcal{K} \subseteq \{1, 2, \ldots, K\}$. The instrumental and direct comparisons in $\mathcal{K}$ are biased by at most $\Gamma_k \geq 1$, $k \in \mathcal{K}$ if: (i) partial exclusion holds for $\mathcal{K}$, and (ii) for each $k \in \mathcal{K}$, treatment assignment is governed by (5.2) with $\gamma_k = \log (\Gamma_k)$ for some unknown $\mathbf{u}_k \in \mathcal{U}$.

If the instrumental and direct comparisons in $\mathcal{K}$ are biased by at most $\Gamma_k \geq 1$, $k \in \mathcal{K}$, then we could perform $|\mathcal{K}|$ separate stratified sensitivity analyses for $|\mathcal{K}|$ tests of $H_0$, with one upper bound $\overline{P}_{k, \Gamma_k}$ on the $P$-value for test $k \in \mathcal{K}$. This bound says: if $H_0$ is true and if the bias in treatment assignment in comparison $k \in \mathcal{K}$ is at most $\exp (\gamma_k) = \Gamma_k$, then the chance that $\overline{P}_{k, \Gamma_k} \leq \alpha$ is at most $\alpha$. Moreover, each bound $\overline{P}_{k, \Gamma_k}$ is sharp, being attained for some $\mathbf{u}_k \in \mathcal{U}$ and some $\gamma_k = \log (\Gamma_k)$.

Definition 10 is the special case of Definition 11 with $\Gamma_k = 1$ for $k \in \mathcal{K}$, so it suffices to consider Definition 11 in formal results.

**5.4.2. Evidence factors**

The discussion in §5.4.1 says that if the comparisons in $\mathcal{K}$ are biased by at most $\Gamma_k \geq 1$, $k \in \mathcal{K}$, then we may obtain $|\mathcal{K}|$ upper bounds $\overline{P}_{k, \Gamma_k}$ on valid $P$-values testing $H_0$, where these $|\mathcal{K}|$ tests make different assumptions about which instruments and comparisons are valid or biased to a limited degree. How are these $|\mathcal{K}|$ analyses related? Are they strongly dependent, merely repeating the same evidence in different forms? Is it nearly a foregone conclusion that the $|\mathcal{K}|$ comparisons will concur? Or are $|\mathcal{K}|$ comparisons nearly statistically independent, so that each comparison provides new evidence? Proposition 2 shows that the $|\mathcal{K}|$ random variables $\overline{P}_{k, \Gamma_k}$ may be treated as if they were statistically independent $P$-values under $H_0$ if the comparisons in $\mathcal{K}$ are biased by at most $\Gamma_k \geq 1$, $k \in \mathcal{K}$.

**Proposition 2** If $H_0$ is true and the comparisons in $\mathcal{K}$ are biased by at most $\Gamma_k \geq 1$, 

94
$k \in \mathcal{K}$, then the $|\mathcal{K}|$ bounds $\overline{p}_{k, \Gamma_k}$ on $P$-values testing $H_0$ are stochastically larger than the uniform distribution of the $|\mathcal{K}|$-dimensional unit cube:

$$\Pr (\overline{p}_{k, \Gamma_k} \leq \alpha_k, \forall k \in \mathcal{K}) \leq \prod_{k \in \mathcal{K}} \alpha_k \text{ for all } 0 \leq \alpha_k \leq 1. \quad (5.3)$$

**Proof.** The proof uses Lemma 4 in Rosenbaum (2011) and runs parallel to the proof of Proposition 3 of that chapter. The stratified structure with instruments in (5.1) is different from the structure in Rosenbaum (2011), but these differences do not affect the proof.

Being stochastically larger than the uniform distribution, as in (5.3), is not quite the same as being statistically independent, but (5.3) is what is needed for hypothesis testing. A variety of methods exist for combining $|\mathcal{K}|$ independent $P$-values that test one hypothesis, resulting in a single $P$-value for the combination. Typically, the combined test statistic is a monotone function of the $|\mathcal{K}|$-dimensional vector of $P$-values, and Lemma 1 of Rosenbaum (2011) shows that such a combination yields a valid combined $P$-value when the components are stochastically larger than the uniform, as in (5.3). Fisher combined independent $P$-values using their product, or equivalently using minus two times the sum of their logs, comparing the latter to the chi-square distribution on $2 \times |\mathcal{K}|$ degrees of freedom. Zaykin et al. (2002) combined independent $P$-values using a truncated product, specifically the product of those $P$-values that are smaller than some truncation point, $\varkappa$, so their method becomes Fisher’s method when $\varkappa = 1$. Hsu et al. (2013) show that the truncated product with $\varkappa = 0.2$ or $\varkappa = 0.1$ often has higher power than Fisher’s method when applied to $P$-value bounds $\overline{p}_{k, \Gamma_k}$ from a sensitivity analysis, because the individual bounds are not uniform on $[0, 1]$ but rather stochastically larger than the uniform.
5.5. Evaluating the performance of the proposed analysis

5.5.1. A model for evaluating performance

The method in §5.4 considers various assumptions that might possibly identify causal effects, but it makes few other assumptions. To evaluate the performance of that method in comparison to other methods, such as two-stage least squares, we consider a specific model for response $R$ in terms of possibly invalid binary instruments $Z_1$ and $Z_2$ and treatment $Z_3$:

$$R = \alpha + \lambda_1 Z_1 + \lambda_2 Z_2 + \beta Z_3 + \epsilon$$

(5.4)

$$\zeta = \nu + \psi_1 Z_1 + \psi_2 Z_2 + \eta \text{ with } E(\epsilon, \eta | Z_1, Z_2) = (0, 0),$$

(5.5)

$$\Pr (Z_3 = 1 | Z_1, Z_2, \eta) = \max \{0, \min (1, \zeta)\}$$

(5.6)

where the bivariate $(\epsilon, \eta)$ are independent and identically distributed given $Z_1$, $Z_2$ with finite variances; see Small (2007) for a related model. We follow Dawid (1979) and write $A \perp \perp B \mid C$ for $A$ is conditionally independent of $B$ given $C$.

If $\epsilon$ and $\eta$ are unrelated, then we do not need instruments, and we could compare treated and control groups directly after adjusting for $Z_1$ and $Z_2$. More precisely, if $\epsilon \perp \perp \eta | (Z_1, Z_2)$, then $\epsilon \perp \perp Z_3 | (Z_1, Z_2)$, and we can draw inferences about $\beta$ using (5.4) alone, adjusting for $(Z_1, Z_2)$, comparing the treated $Z_3 = 1$ and control $Z_3 = 0$ groups directly. For instance, inference about $\beta$ could be based on a least squares regression in (5.4), ignoring (5.5) and (5.6). Alternatively, inference about $\beta$ could be based on the direct comparison of treated and control groups stratified for $Z_1$ and $Z_2$; that is, (iii) with $\gamma_3 = 0$ in §5.2 or step $k = 3$ in §5.3.

If $\epsilon$ and $\eta$ were dependent given $(Z_1, Z_2)$, but $\lambda_1 = \lambda_2 = 0$ with $\psi_1 \neq 0$ and $\psi_2 \neq 0$, then $Z_1$, $Z_2$ would be instruments for $Z_3$. For instance, inference about $\beta$ could be based on two-stage least squares. Also, factors $k = 1$ and $k = 2$ in §5.3 would each provide valid inferences about $\beta$ with $\gamma_1 = 0$ and $\gamma_2 = 0$. If $\epsilon$ and $\eta$ were dependent but either $\lambda_1 \neq 0$ or
If $\epsilon$ and $\eta$ were dependent given $(Z_1, Z_2)$, but $\lambda_2 = 0$ with $\psi_2 \neq 0$, then $Z_2$ would be a valid instrument for $Z_3$ after adjustment for $Z_1$. For instance, after stratifying for $Z_1$, factor $k = 2$ in §5.3 would provide valid inferences about $\beta$ with $\gamma_2 = 0$. However, even factor $k = 2$ would be invalid if $\epsilon$ and $\eta$ were dependent but $\lambda_2 \neq 0$.

5.5.2. Details of the model for numerical results

To obtain some numerical results, we further specified the distributions in (5.4)-(5.6). Parameters were either fixed or variable, with some fixed parameters chosen to resemble actual distributions in the example in §5.2. In particular, we set $\Pr(Z_1 = 1) = 0.33$, $\Pr(Z_2 = 1) = 0.40$ as fixed parameters, but we measured dependence between $Z_1$ and $Z_2$ by a variable parameter $\delta = \Pr(Z_2 = 1|Z_1 = 1) - \Pr(Z_2 = 1|Z_1 = 0)$, with either $\delta = 0$ for independence or $\delta = 0.14$ for dependence. Given $(Z_1, Z_2)$, the two errors $(\epsilon, \eta)$ were bivariate Normal with zero expectations, variable correlation $\rho$, and fixed variances 1 and 0.06. In the simulation, to resemble §5.2, we set $\psi_1 = 0.20$ and $\psi_1 = 0.25$, but in later calculations we varied these parameters to include weak instruments. We set $\nu = 0$, varying $\lambda_1$ and $\lambda_2$. In the simulation, the sample size is $N = 4450$, with $I = 178$ strata, $i = 1, \ldots, I$, of size $n_i = 25$, as in the example, whereas calculations of design sensitivity let $I \to \infty$ with $n_i = 25$.

5.5.3. Simulation of the probability of finding an effect when there is none

The simulation concerns the validity of various tests of a true null hypothesis that $H_0 : \beta = \beta_0$. The simulation evaluates the size of a test that aspires to have level 0.05, so the test succeeds if its size is at most 0.05 and fails if it rejects the true null hypothesis at a rate above 0.05. Theory tells us whether a test is valid (V), with size at most equal to level, or biased (B), with size sometimes above level, and the judgement of theory is indicated by a B or V in Table 9. The simulated sizes do agree with theory, but they also provide a
Table 9: Simulated probability of rejecting, at the 0.05 level, the true null hypothesis \( H_0 : \beta = \beta_0 \) using two-stage least squares (TSLS), (i) using the first binary instrument alone, (ii) using the second binary instrument stratifying for the first instrument, (iii) using a direct comparison of treated and control groups stratifying for both instruments. Because \( H_0 \) is true, every rejection is a mistake. If the test is valid (V), then the probability of rejection is 0.05, but if the test is biased (B) then the probability of rejection is inflated. The evidence-factor analysis fails to provide a warning if all three factors, (i)-(iii), are biased (B), which occurs in three cases, 8, 15 and 16. TSLS fails in 12 cases.

<table>
<thead>
<tr>
<th>Case</th>
<th>Parameters</th>
<th>Valid (V) or asymptotically biased (B)</th>
<th>Probability of rejecting ( H_0 : \beta = \beta_0 )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( \lambda_1 )</td>
<td>( \lambda_2 )</td>
<td>( \rho )</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>0.10</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>0.10</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>0.10</td>
<td>0.10</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0.82</td>
</tr>
<tr>
<td>6</td>
<td>0.10</td>
<td>0</td>
<td>0.82</td>
</tr>
<tr>
<td>7</td>
<td>0</td>
<td>0.10</td>
<td>0.82</td>
</tr>
<tr>
<td>8</td>
<td>0.10</td>
<td>0.10</td>
<td>0.82</td>
</tr>
<tr>
<td>9</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>10</td>
<td>0.10</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>11</td>
<td>0</td>
<td>0.10</td>
<td>0</td>
</tr>
<tr>
<td>12</td>
<td>0.10</td>
<td>0.10</td>
<td>0</td>
</tr>
<tr>
<td>13</td>
<td>0</td>
<td>0</td>
<td>0.82</td>
</tr>
<tr>
<td>14</td>
<td>0.10</td>
<td>0</td>
<td>0.82</td>
</tr>
<tr>
<td>15</td>
<td>0</td>
<td>0.10</td>
<td>0.82</td>
</tr>
<tr>
<td>16</td>
<td>0.10</td>
<td>0.10</td>
<td>0.82</td>
</tr>
</tbody>
</table>

The quantitative dimension that goes beyond the dichotomy. Table 9 compares four methods, namely two-stage least squares using both \((Z_1, Z_2)\), and the three evidence factors that use \(Z_j\) adjusting for \(Z_{j-1}, \ldots, Z_1\).

Table 9 shows the simulation results, where each sampling situation was replicated 10,000 times, so that a binomial proportion has a standard error of at most \(\sqrt{0.25/10000} = 0.005\).

Table 9 has sixteen sampling situations. In case 5, for example, all three tests of \(H_0\) based on instruments have the correct level, but the direct comparison is all but certain to falsely reject \(H_0\).

All four tests are valid only in cases 1 and 9, in which both instruments and the direct
comparison are valid. In total, two-stage least squares is valid in only four cases, 1, 5, 9, and 13, in which both instruments are valid.

Suppose that an investigator rejected $H_0$ only if all three evidence factors concur in rejecting $H_0$. Used in this way, the evidence factor analysis fails to provide a warning about biased comparisons only in cases 8, 15 and 16; it is valid in 13/16 cases. More formally, suppose at least one of the three factors is valid. Under this supposition, following Berger’s (1982) reasoning about intersection-union tests, if we reject $H_0$ only when all three factors reject $H_0$, then we would falsely reject $H_0$ with probability at most 0.05.

Alternatively, a weaker standard in §5.6.4 uses the idea and method of partial conjunction from Benjamini and Heller (2008), saying that the evidence factors partially concur if at least two of them reject $H_0$. This weaker standard would fail, for instance, in case 4 where two factors are likely to falsely reject $H_0$; however, it provides protection whenever there is a single B in columns (i)-(iii). For instance, it provides protection in cases 2 and 3 with one invalid instrument, and in case 5 where both instruments are valid but the direct comparison is not.

To require evidence factors to concur is to require agreement among several nearly independent analyses that are valid under different assumptions. Table 9 shows that this approach is not infallible, but it does offer substantially more protection than opting for any single analysis, say two-stage least squares or the direct treatment-control comparison (iii).

5.5.4. Design sensitivity

Section 5.5.3 examined the protection afforded by evidence factors against falsely rejecting a true $H_0$. We now consider testing a false null hypothesis, one that we hope to reject for valid reasons, that is, in the subset of valid analyses. Specifically, we test $H_0 : \beta = 0$ when in fact $\beta = 0.5$. So, in each sampling situation, we delete the biased analyses, as we may reject here because of bias, and we consider our prospects for rejecting $H_0$ in the remaining valid analyses.
Fix a sampling situation and let the sample size increase, \( I \to \infty \). For each fixed \( \Gamma \), the power of a sensitivity analysis tends either to 1 or to 0 as \( I \to \infty \), depending upon the value of \( \Gamma \). The transition point, \( \tilde{\Gamma} \), is called the design sensitivity: the power tends to 1 if \( \Gamma < \tilde{\Gamma} \) or to 0 if \( \Gamma > \tilde{\Gamma} \). In that sense, \( \tilde{\Gamma} \) is the limiting sensitivity to unmeasured bias when sampling variability has been eliminated by a sufficient increase in sample size.

Table 10 shows design sensitivities. Unlike Table 9, in Table 10 the instruments are sometimes weak, sometimes strong, and sometimes one is weak when the other is strong. It is known from theory that an analysis that uses a weak instrument is invariably sensitive to small biases, its design sensitivity \( \tilde{\Gamma} \) being barely larger than 1; see Small and Rosenbaum (2008).

Table 10 reminds us of a couple of basic quantitative facts. First, when there is a substantial treatment effect and no unmeasured bias, a direct comparison of treated and control groups may be insensitive to quite large biases. In a matched pair, a bias of \( \Gamma = 2.5 \) could be produced by an unobserved covariate \( u \) that increases the odds of treatment by a factor of 4 and increases the odds of a positive pair difference in outcomes by a factor of 6; see Rosenbaum and Silber (2009). Even if there is reason to worry that a direct comparison might be slightly biased, we may discover that it would have to be very biased to change the study’s qualitative conclusion.

In contrast, the instrumental variable analyses in Table 10 are all sensitive to smaller biases. A bias of \( \Gamma = 1.25 \) is not trivially small: in a matched pair, it could be produced by an unobserved covariate that doubled the odds of treatment and doubled the odds of a positive pair difference in responses. In Table 10, design sensitivities of \( \tilde{\Gamma} \approx 1.2 \) occur with strong instruments or \( \tilde{\Gamma} \approx 1.06 \) with weak instruments. A weak instrument has to be almost flawless to be convincing.
Table 10: Design sensitivities $\bar{F}$ for valid analyses. For biased analyses, a “B” is given in place of a design sensitivity.

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Two strong instruments $\psi_1 = 0.20, \psi_2 = 0.25$</th>
<th>Two weak instruments $\psi_1 = 0.09, \psi_2 = 0.09$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\lambda_1$</td>
<td>$\lambda_2$</td>
<td>$\rho$</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0.10</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0.10</td>
<td>0.10</td>
<td>0</td>
</tr>
<tr>
<td>0.10</td>
<td>0.10</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0.10</td>
<td>0.14</td>
</tr>
<tr>
<td>0.10</td>
<td>0.10</td>
<td>0.14</td>
</tr>
<tr>
<td>0.10</td>
<td>0.10</td>
<td>0.14</td>
</tr>
<tr>
<td>0.10</td>
<td>0.10</td>
<td>0.14</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Parameters</th>
<th>$Z_1$ weak, $Z_2$ strong $\psi_1 = 0.09, \psi_2 = 0.25$</th>
<th>$Z_1$ strong, $Z_2$ weak $\psi_1 = 0.20, \psi_2 = .09$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\lambda_1$</td>
<td>$\lambda_2$</td>
<td>$\rho$</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0.10</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0.10</td>
<td>0.10</td>
<td>0</td>
</tr>
<tr>
<td>0.10</td>
<td>0.10</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0.10</td>
<td>0.14</td>
</tr>
<tr>
<td>0.10</td>
<td>0.10</td>
<td>0.14</td>
</tr>
<tr>
<td>0.10</td>
<td>0.10</td>
<td>0.14</td>
</tr>
<tr>
<td>0</td>
<td>0.10</td>
<td>0.14</td>
</tr>
<tr>
<td>0</td>
<td>0.10</td>
<td>0.14</td>
</tr>
<tr>
<td>0.10</td>
<td>0.10</td>
<td>0.14</td>
</tr>
<tr>
<td>0.10</td>
<td>0.10</td>
<td>0.14</td>
</tr>
<tr>
<td>0</td>
<td>0.10</td>
<td>0.14</td>
</tr>
<tr>
<td>0</td>
<td>0.10</td>
<td>0.14</td>
</tr>
<tr>
<td>0.10</td>
<td>0.10</td>
<td>0.14</td>
</tr>
<tr>
<td>0</td>
<td>0.10</td>
<td>0.14</td>
</tr>
<tr>
<td>0.10</td>
<td>0.10</td>
<td>0.14</td>
</tr>
</tbody>
</table>
5.6. Effects of Catholic versus public high schools

5.6.1. Adjustments for observed covariates

In our examination of income from wages and salary for men in 1974, a preliminary step is to adjust for observed covariates. A simple analysis uses 178 strata defined by covariates, and a second analysis combines these 178 strata with a robust covariance adjustment. Following Kim (2011), we adjust for an IQ score prior to high school, father’s and mother’s education, parent’s income, father’s occupation score and occupational prestige score. Missing values in covariates were handled using the tactic in Rosenbaum and Rubin (1984, Appendix) in which treated subjects are compared to controls with a similar pattern of missing data.

The strata were built using the heuristic in the blockingChallenge package in R, where details may be found. The heuristic samples 178 students at random, then uses optimal matching to match 24 other students to each of the initial 178 students, making 178 blocks of size 25. The matching minimizes a robust Mahalanobis distance based on the observed covariates. The one student in each block most distant from the remaining 24 is separated, and optimal matching is used again to pair these 178 individual students with 178 blocks of size 24. This process is repeated until no further changes are produced. The whole process was done 250 times, with 250 different random samples of 178 students, and we used the best of the 250 stratifications, that is, the one with the smallest total within-block distance.

The package blockingChallenge has the name it does because it is an invitation to produce a better heuristic or approximation algorithm for minimum distance stratification.

Figure 1 depicts the effects of the stratification. The dashed line is the density estimate for the marginal distribution of \( 178 \times 25 = 4450 \) values of the covariate. The solid line shows the density estimate after removing the block means: the 178 block means were subtracted, and the grand mean was added back. Each plot is labeled with the F-ratio for a one-way anova with 178 groups defined by the blocks. For instance, the F-ratio for IQ is 225.6, so the variation between the 178 blocks is substantial.
The covariance adjustment used the method in Rosenbaum (2002). The outcome, income from wages and salary in 1974, was regressed using M-estimation on indicators for the 178 strata plus the covariates themselves, and the residuals became the outcome to be analyzed by the method in §5.4. Importantly, this regression used the strata and covariates, but not the treatment variables in Table 8. Use of this form of covariance adjustment with an instrument is discussed and illustrated in Rosenbaum (2002).

Table 11: Three evidence factors and their combination using the truncated product, with and without covariance adjustment. The case $\Gamma = 1$ assumes comparisons are flawless, three stratified randomized experiments. The table shows one equivalent amplification of each $\Gamma > 1$. The table displays upper bounds on one-sided $P$-values testing the null hypothesis that Catholic schooling raises wages by at most $\beta$ dollars in the presence of a bias of at most $\Gamma$. As the median annual wage was $14000$, a $500$ increase is about 3.6%.

<table>
<thead>
<tr>
<th>Sensitivity Parameter</th>
<th>Equivalent Amplification $(\Lambda, \Delta)$</th>
<th>3 Independent Factors</th>
<th>Stratified analysis</th>
<th>Combined</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\Gamma$</td>
<td></td>
<td></td>
<td>Stratified analysis</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$H_0 : \beta = 0$</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>(1, 1)</td>
<td>0.0000</td>
<td>0.0041</td>
<td>0.0082</td>
</tr>
<tr>
<td>1.1</td>
<td>(1.4, 1.8)</td>
<td>0.0000</td>
<td>0.0835</td>
<td>0.0422</td>
</tr>
<tr>
<td>1.2</td>
<td>(1.75, 2)</td>
<td>0.0004</td>
<td>0.4095</td>
<td>0.1331</td>
</tr>
<tr>
<td>1.25</td>
<td>(2, 2)</td>
<td>0.0023</td>
<td>0.6225</td>
<td>0.2049</td>
</tr>
<tr>
<td>$\Gamma$</td>
<td></td>
<td></td>
<td>Stratified + covariance adjustment</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$H_0 : \beta = 0$</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>(1, 1)</td>
<td>0.0000</td>
<td>0.0065</td>
<td>0.0149</td>
</tr>
<tr>
<td>1.1</td>
<td>(1.4, 1.8)</td>
<td>0.0001</td>
<td>0.1115</td>
<td>0.0667</td>
</tr>
<tr>
<td>1.2</td>
<td>(1.75, 2)</td>
<td>0.0048</td>
<td>0.4738</td>
<td>0.1876</td>
</tr>
<tr>
<td>1.25</td>
<td>(2, 2)</td>
<td>0.0182</td>
<td>0.6827</td>
<td>0.2747</td>
</tr>
<tr>
<td>$\Gamma$</td>
<td></td>
<td></td>
<td>$H_0 : \beta = 500$</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>(1, 1)</td>
<td>0.0000</td>
<td>0.0394</td>
<td>0.2013</td>
</tr>
<tr>
<td>1.1</td>
<td>(1.4, 1.8)</td>
<td>0.0005</td>
<td>0.3105</td>
<td>0.4345</td>
</tr>
<tr>
<td>1.2</td>
<td>(1.75, 2)</td>
<td>0.0128</td>
<td>0.7451</td>
<td>0.6735</td>
</tr>
<tr>
<td>$\Gamma$</td>
<td></td>
<td></td>
<td>$H_0 : \beta = 1000$</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>(1, 1)</td>
<td>0.0000</td>
<td>0.1304</td>
<td>0.6975</td>
</tr>
<tr>
<td>1.1</td>
<td>(1.4, 1.8)</td>
<td>0.0016</td>
<td>0.5550</td>
<td>0.8826</td>
</tr>
<tr>
<td>1.2</td>
<td>(1.75, 2)</td>
<td>0.0303</td>
<td>0.9018</td>
<td>0.9643</td>
</tr>
</tbody>
</table>

5.6.2. Naive analysis: each comparison is either flawless or useless

Table 11 performs the analyses from §5.4 on the wage data, testing three hypotheses, namely that Catholic schooling does not increase wages, $H_0 : \beta \leq 0$, that it increases wages by at
most $500, $H_0 : \beta \leq 500$, and that it increases wages by at most $1000, H_0 : \beta \leq 1000$.

As the median annual wage in 1974 for these men was $14000, an increase of $500 is about
3.6%. Two analyses are performed for $H_0 : \beta \leq 0$, namely a stratified analysis, and a
stratified analysis on residuals from covariance adjustment. Because there is no reason to
prefer the merely stratified analysis, the latter analysis is presented in greater detail.

The current section assumes that each of the three comparisons is essentially a randomized
experiment, once adjustments have been made for observed covariates. This is the situation
with $\Gamma = 1$ in Table 11. The case of $\Gamma > 1$ is discussed in §5.6.3.

In both the stratified analysis and the covariance adjusted stratified analysis, each of the
three comparisons rejects the null hypothesis of no effect of Catholic schools on wages,
so the three evidence factors concur. These three factors depend upon very different
assumptions, and they would be nearly statistically independent were the null hypothesis
true, so it is news that the three analyses concur. When the three analyses are pooled using
the truncated product of $P$-values, with the default truncation of 0.2, the resulting $P$-value
is extremely small.

When testing the hypothesis that the effect is at most $500$, the situation is quite different.
Two analyses reject $500$ as too small, but the remaining factor does not concur. The pooled
analysis is significant because of the urban/rural comparison; remove that, and the pooled
$P$-value from the two remaining factors is 0.103. A similar pattern is seen when testing
that the effect is at most $1000$.

5.6.3. Sensitivity analysis: allowing for small or moderate imperfections in each analysis

How might small unmeasured biases alter the analyses in §5.6.2? Table 11 considers biases
of $\Gamma = 1.1$, 1.2, and 1.25. The parameter $\Gamma$ has the same meaning in matched pairs and in
strata, but it is easiest to understand the paired case; see Rosenbaum (2017, Table 9.1). If
we paired people based on covariates and flipped a fair coin to assign one person in the pair
to live in an urban area, or to be Catholic, or to attend Catholic school, then each person
in the pair would have probability 1/2 of each of these assignments. If \( \Gamma = 1.2 \), then the same probability is somewhere in the interval [0.455, 0.545] rather than 1/2. Equivalently, \( \Gamma \) may be given an equivalent two-parameter interpretation: in a pair, \( \Gamma = 1.2 \) is the same as an unobserved covariate than increases the odds of treatment by a factor of \( \Lambda = 1.75 \) and doubles the odds, \( \Delta = 2 \), of a positive pair difference in wages. A bias of \( \Gamma = 1.1 \) is small, and would be hard to rule out based on a priori considerations in most observational studies. A bias of \( \Gamma = 1.25 \) is neither large nor trivially small: it corresponds with an unobserved covariate that doubles the odds of treatment and doubles the odds of a positive pair difference in wages.

In Table 11, the pooled test of no effect of Catholic school using stratification and covariance is insensitive to a bias \( \Gamma = 1.2 \); however, this is entirely due to the contribution of the urban/rural instrument. Without the urban/rural instrument, the pooled test of no effect using the other two factors is sensitive at \( \Gamma = 1.2 \).

So, the analysis depends rather heavily on the validity of urban/rural as an instrument. The instrumental variable analysis notes higher wages for students from urban areas, and attributes that difference in wages to a higher frequency of Catholic schooling in urban areas. In fact, that attribution is somewhat suspect here. Among non-Catholics attending public school, median wages were higher in urban areas, a median of $15000 in urban areas versus $13000 in rural areas. Among Catholics attending public schools, median wages were higher in urban areas, a median of $14000 in urban areas versus $13400 in rural areas. It is a source of concern that the analysis depends so heavily on the urban/rural instrument, as it is plausible that wages are higher for students from urban areas for reasons other than Catholic schooling.

5.6.4. Partial conjunction

As just noted, the combined analyses in Table 11 lean heavily on the validity of urban/rural as an instrument. Is it possible to quantify the degree to which a combined analysis depends
upon one of its components? How large can $\Gamma$ be while still securing concurrence in rejecting $H_0$ by at least two components?

Partial conjunction hypotheses ask for concurrence among at least $K$ sources of evidence, $1 < K < K$, without specifying in advance which $K$ sources will concur. In Table 11, $K = 3$ so the only possible value of $K$ is 2. The partial conjunction null hypothesis asserts that at most $K - 1$ null hypotheses are false, so rejection of that hypothesis entails at least $K$ null hypotheses are false. In Table 11, we seek strong evidence that at least $K = 2$ factors concur in rejecting $H_0$. The original interest in partial conjunction hypotheses arose in neuroimaging experiments, and methods have been studied systematically by Benjamini and Heller (2008) and Wang and Owen (2018). Applying results in these articles to Proposition 2, we may reject at level $\alpha$ the $K$ partial conjunction hypothesis in the presence of biases of at most $\Gamma_k$, $k = 1, \ldots, K$, if the $P$-value determined by the truncated product is $\leq \alpha$ when computed from $K - K + 1 = 3 - 2 + 1 = 2$ largest $P_{k, \Gamma_k}$. In Table 11, the smallest $K - 1 = 1$ smallest $P_{k, \Gamma_k}$ is always from $k = 1$ for the urban/rural comparison; however, this method acknowledges that we could not know that prior to examining the data.

Consider partial conjunction testing of $H_0 : \beta = 0$ using both stratification and covariance adjustment in Table 11. In randomization tests, $\Gamma_1 = \Gamma_2 = \Gamma_3 = 1$, applying the truncated product to $P_{2, \Gamma_2} = 0.0065$ and $P_{3, \Gamma_3} = 0.0149$ yields a $P$-value of 0.00084, so at least two factors concur in rejecting $H_0$. At $\Gamma_1 = \Gamma_2 = \Gamma_3 = 1.1$, the two bounds, $P_{2, \Gamma_2} = 0.1115$ and $P_{3, \Gamma_3} = 0.0667$, combine to yield a $P$-value of 0.0319; however, $\Gamma_1 = \Gamma_2 = \Gamma_3 = 1.2$, the combined $P$-value is 0.34. In short, at least two factors concur in rejecting $H_0$ if the unmeasured bias is quite small, $\Gamma_1 = \Gamma_2 = \Gamma_3 = 1.1$, but for larger biases the analysis rejection depends entirely on the validity of the urban/rural instrument.

5.7. Discussion

Conventional analyses with two or more instruments, such as two-stage least squares, assume all instruments are jointly flawless, and ignore the direct comparison of treated and control
groups. In contrast, our proposed analysis assumes less and reveals more. With $K - 1$ instruments, we produce $K$ essentially independent comparisons that each make very different assumptions for the validity of different comparisons, successively changing the role of each instrument from instrument to covariate. To a considerable extent, neither identical assumptions nor chance can explain concurrence among these $K$ analyses, whereas an actual causal effect could explain concurrence. Each analysis is subjected to a sensitivity analysis that quantitatively evaluates the gradual failure of the assumptions upon which that one analysis depends. A partial conjunction analysis asks about the evidence that remains when the quantitatively most compelling analyses are set aside.

Because instruments, particularly weak instruments, are greatly unsettled by the slightest flaw, theory suggests that the least sensitive finding — the finding with the largest design sensitivity $\tilde{\Gamma}$ — is expected to come from the direct comparison of treated and control groups when, indeed, there is a causal effect without bias. This suggests, the direct comparison should be one of the $K$ factors considered and displayed. True, the direct comparison may be the most biased comparison, so we might tolerate sensitivity to smaller biases in an instrument than in a direct comparison.
CHAPTER 6: Assessment of the extent of corroboration of an elaborate theory of a causal hypothesis using partial conjunctions of evidence factors

6.1. Introduction

6.1.1. An elaborate theory of a causal effect and evidence factors

Fisher’s response to the question “what can be done in observational studies to clarify the step from association to causation?” was: “Make your theories elaborate” (Cochran, 1965). Cochran explains this response by stating that to clarify the step from association to causation one should envision as many different consequences as possible of the causal hypothesis under investigation and design studies which are able to scrutinize these consequences. In parallel to Cochran’s interpretation of Fisher’s response, Popper (1934, 1972), through arguments of classical logic, emphasizes the importance to scientific progress for a hypothesis to have a higher ‘degree of testability’. By degree of testability, Popper means the amount of falsifiable ‘basic statements’ the theory generates. “If we look for confirmations”, Popper (1963) writes, “It is easy to obtain confirmations ...for nearly every theory”, while “[e]very genuine test of a theory is an attempt to falsify it, or to refute it. Testability is falsifiability.”

The motivating example of this chapter, discussed in detail in §6.2, considers the causal hypothesis that exposure to lead of a parent at the workplace causes high level of lead in the blood of a child at home. To test this causal hypothesis, Morton et al. (1982) established the following elaborate theory (Rosenbaum, 2005): (a) children of parents who were occupationally exposed to lead will have higher lead levels in the blood than otherwise similar control children; (b) among children of parents occupationally exposed to lead, children of parents with higher occupational lead exposure will have higher lead levels than otherwise similar children of parents with lower occupational lead exposure; and (c) among children of parents occupationally exposed to lead, children whose parents practiced poorer hygiene before leaving work will have higher lead levels than otherwise similar children whose parents practiced better hygiene. We are interested in the question: what is the extent of
corroboration of this theory provided by the data? Popper (1972), in the addendum to his final chapter of *The Logic of Scientific Discovery*, writes, “I tried to make clear that by the degree of corroboration of a theory I mean a brief report that summarizes the way in which the theory has stood up to tests, and how severe these tests were.” In practice, the degree of corroboration of an elaborate theory has been evaluated by reporting what fraction of test of predictions of the elaborate theory have p-values < 0.05 (where rejecting the null supports the elaborate theory); see, e.g., Centerwall (1989) or Wong, Cook and Steiner (2015).

There are two problems with just counting the fraction of p-values less than 0.05 for assessing degree of corroboration of an elaborate theory. First, if the tests are dependent, then multiple tests rejecting may not be providing much more evidence than one test rejecting. Second, counting the fraction of p-values less than 0.05 is not an efficient combination of the evidence. For example, if two independent tests of the same null hypothesis both have p-values 0.06, this is strong evidence against the null by Fisher’s method of combining independent tests (Fisher, 1932), the p-value for Fisher’s combined test is 0.02.

An additional problem with the current practice for assessing the degree of corroboration for an elaborate theory is that the p-value computed for each test of the elaborate theory assumes no unmeasured confounding. In most observational studies, unmeasured confounding is a concern, and we would not find convincing an inference that was valid with no unmeasured confounding but invalid with a little bit of unmeasured confounding. A sensitivity analysis examines how much bias from unmeasured confounding could change the conclusions of a study that assumed no unmeasured confounding (Cornfield et al., 1959; Rosenbaum, 1987; Hosman et al., 2010; Keele and Minozzi, 2013; Stuart et al., 2013; Ding and Vanderweele, 2016; Fogarty and Hasegawa, 2018).

We develop a method for assessing the extent of corroboration of an elaborate theory that overcomes the three shortcomings we identified above of the current p-value counting approach. Our method involves three aspects: (i) we decompose the test of the elaborate the-
ory into evidence factors, pieces that are affected by different biases and statistically near independent (Rosenbaum, 2011, 2017; Zubizarreta et al., 2012) (the additional requirement of different biases in each test increases robustness of the analysis against multiple potential sources of biases); (ii) we assess the extent of corroboration in a way that combines the information from different tests efficiently and furthermore we use partial conjunction tests (Benjamini and Heller, 2008; Benjamini, Heller, and Yekutieli, 2009); and (iii) we test the evidence factors using sensitivity analysis methods that allow for specified amounts of unmeasured confounding. The novel contributions of the chapter are the following: (a) we provide a systematic approach to decomposing an elaborate theory into evidence factors; (b) as a way to test for partial corroboration of the elaborate theory, we introduce partial conjunction tests (partial conjunction tests have been previously developed for the purpose of inference in neuroimaging experiments by Benjamini and Heller, 2008); (c) we develop a sensitivity analysis method for carrying out (a) and (b) that allows for a specified degree of unmeasured confounding; (d) we show that the method developed for (c) controls for the overall familywise error rate in the multi-parameter sensitivity analysis; and (e) for the method for (c), which involves combining sensitivity analyses for each of the evidence factors, we find the asymptotically optimal such combining method.

6.1.2. Sir Karl Popper and degree of corroboration

The term ‘degree of corroboration’ was introduced by Popper in response to an inattentive translation, ‘degree of confirmation’, of his original phrase ‘Grad der Bewährung’. Two decades after Logik der Forschung, in three Br. J. Philos. Sci. notes (vol. 5, pp. 143–149, 1954; vol. 7, pp. 350–353, 1957; and vol. 8, pp. 294–302, 1958) Popper came up with a definition of degree of confirmation or degree of corroboration. In these notes, his motivation was rather different. He first attempted to show that, in the sense it is to be used in science, degree of corroboration or acceptability of a theory cannot be a probability. After showing this, he suggested a definition of the degree to which a statement x is confirmed by a statement y which he named the degree of confirmation of x by y. This definition was
based on a list of desiderata he had put down for such a quantity. This definition may serve its purpose, but does not serve ours. First, such a definition depends on a background probability measure appropriately defined on first-order languages, and computations under this probability measure have not been well developed for statistical practice (Popper, 1954; Crupi, Chater and Tentori, 2013). Second, it is still an unsettled debate whether such a quantity is an adequate measure of corroboration (Rowbottom, 2013; Sprenger, 2018). Finally, this definition attempts to answer a very different question than ours. We are interested in the investigation of a causal hypothesis in an observational study and how best to make inferences about it from a frequentist perspective, whereas Popper attempted to define a quantity which would replace the p-value in investigation of a scientific theory.

6.1.3. Outline of the chapter

The chapter is organized as follows. We discuss our motivating example in §6.2. Here we briefly recall the original study. The notation for our method is introduced in §6.3.1. Section 6.3.2 recalls the treatment assignment models for the observed data. A brief review of the testing procedures and their sensitivity analysis is given in §6.3.3. The decomposition of the tests into evidence factors is established in §6.4. Our main method is developed in §6.5. In particular, Proposition 13 defines the (maximum) p-values for tests of partial conjunction of the hypotheses. Using these p-values we get tests of all the partial conjunctions of the hypotheses for any given value of the sensitivity parameters. Theorem 14 and its corollaries show that the familywise error rate is controlled in our multi-parameter sensitivity analysis, with a range of values of the bias parameter, for the tests of the collection of all the partial conjunctions of the hypotheses. Section 6.6 compares the methods of testing the elaborate theory in their performance in sensitivity analysis. Section 6.6.2 finds asymptotically optimal methods in sensitivity analysis for tests of partial conjunctions of the hypotheses for elaborate theories. In §6.6.3, a simulation study is used for comparison of various methods in their power of sensitivity analysis. The simulation show that methods that pool evidence from the various evidence factors are favorable over methods that look at the individual...
tests who lose power when looking at fractions of the elaborate theory. Results of the study in §6.2 are in §6.7 and the chapter ends with a short conclusion in §6.8.

6.2. Lead absorption study of Morton et al.

6.2.1. The elaborate theory and the analysis

Morton et al. (1982) studied the effect on children of a parent’s occupational exposure to lead. Does exposure of a parent, who works in a battery manufacturing plant (in Oklahoma), to lead at the workplace cause an increase in lead level in the blood of a child in the household? The causal hypothesis is that an employee who is exposed to lead at the workplace carries lead dust back to the household and causes the child to have a higher lead level. To study their elaborate theory, given in §6.1.1, they collected data on 33 matched pairs, with one exposed child and one control child forming a pair. Data were collected on the lead level in the blood of the children; on the lead exposure levels, at the workplace, of the parents of the exposed children — categorized as high, medium, and low; and on hygiene practices of the parents of the exposed children before leaving work — categorized as good, moderately good, and poor.

A multitude of tests were carried out to see if the observed data are consistent with various pieces of the elaborate theory. They found a significantly higher lead level in exposed children compared to their controls. Exposed children of parents with higher lead exposure seemed to have higher lead levels, and parent’s better hygiene practices seemed to indicate a lower lead level in the blood of the children. Focus was not on the separate pieces of the analyses but on the fact that there was a tendency of the evidence to converge to the same direction of confirming the elaborate theory. Although not all the tests corroborated the elaborate theory, e.g., in comparing the exposed children depending on their parent’s lead exposure level, ‘the medium exposure group was not significantly different from the low exposure group,’ the concluding remark of the authors was that the study ‘provides additional confirmation that increased risk of lead absorption occurs in children of employees in a lead-related industry[.]’ Clearly, the strategy was of a multiplist (Reynolds and West,
— several pieces of evidence seeming to converge in favor of the causal hypothesis has been taken as a confirmation of the hypothesis. We will develop a more quantitative approach to summarizing the evidence about the elaborate theory from the study.

6.2.2. *Is there evidence for a causal effect on children of occupational exposure to lead?*

Wilcoxon’s signed rank test for a higher lead level in the blood for exposed child compared to its control has a p-value $P_1 = 6.96 \cdot 10^{-5}$. Among the exposed children, the p-value in comparing high or moderate lead exposure at workplace for the parent versus a low exposure, using Wilcoxon’s rank sum test, is $P_2 = 3.81 \cdot 10^{-3}$. A comparison of exposed children with high lead exposure level of the parents to medium lead exposure level of the parents is $P_3 = 9.59 \cdot 10^{-2}$. Of these three comparisons, the first one tests part (a) of the elaborate theory, the latter two are tests for part (b) of the elaborate theory. For part (c), consider exposed children from families with parent exposed to high level of lead. The p-value is $P_4 = 9.44 \cdot 10^{-3}$ when comparing poor hygiene practice versus a good or moderately good hygiene practice, and the p-value is $P_5 = 0.42$ in comparing a moderately good to a good hygiene practice. Note that for each test, a prediction of a true causal hypothesis is set up as an alternative hypothesis.

If we ask for evidence that all pieces of the elaborate theory are true, we would look at the maximum of those five p-values, which is 0.42. However, if were to pool all the p-values using Fisher’s method — which will be shown using Theorem 12 gives a valid p-value — the pooled p-value is $1.41 \cdot 10^{-6}$, evidence in support of the hypothesis that at least one part of the elaborate theory is true. These are two drastically different numbers — neither suffices for our requirement of representing the extent of corroboration of the elaborate theory offered by the study. If we use the Holm-Bonferroni procedure, it would say that, at level 0.05, there is evidence to reject three out of the five tests, since $(5 + 1 - 3)P_{(3)} = 0.02832 < 0.05$ and $(5 + 1 - 4)P_{(4)} = 0.191846 > 0.05$ (Holm, 1979). We provide the results from our method in §6.7. Our method, which we will now present, looks at the partial conjunction of the tests in combination with a sensitivity analysis.
6.3. Matched pair design with multiple treatments across pairs

6.3.1. Notation: K treatments in I pairs

There are I pairs of units matched on their observed covariates. Let \( i, j \), for \( j = 1, 2 \), index the units in pair \( i, i = 1, \ldots, I \). The observed covariates for unit \( i,j \) are \( x_{ij}, x_{i1} = x_{i2} \) in each pair. Let \( Z_{ij}^{(1)} \) be the indicator of exposure to treatment 1 for unit \( i,j \). In each pair there is one unit with treatment 1 and the other unit is not exposed to that treatment; so \( Z_{i1}^{(1)} + Z_{i2}^{(1)} = 1 \). Each unit is further exposed to treatments 2, \ldots, K. We denote by \( Z_{ij}^{(k)} \) the exposure status to treatment \( k \) for \( i,j \).

In the lead absorption study of §6.2, the first treatment, treatment 1, was employment of a parent in a battery manufacturing plant in Oklahoma. For an exposed child the subsequent treatments were based on parent’s potential occupational exposure to lead — high or medium vs. low, treatment 2 and high vs. medium or low, treatment 3 — and further based on hygiene level of the parent — good or moderately good vs. poor, and good vs. moderately good or poor, treatment 4 and treatment 5 respectively. So, \( I = 33 \) and \( K = 5 \). Morton et al. collected data on occupation level of lead exposure and hygiene practice, only for the individuals exposed to treatment 1. Thus, the data for \( Z_{ij}^{(2)}, \ldots, Z_{ij}^{(5)} \) were not available when \( Z_{ij}^{(1)} = 0 \). This does not hinder our analysis. As will become clear in our methodological development, the effect of treatment 2 will be analyzed only after conditioning on \( Z_{ij} = 1 \). Similarly, the effect of treatment 3 will be assessed only for exposed child with father exposed to high or medium level of occupational lead exposure.

Let \( Z_{ijk} = (Z_{ij}^{(1)}, \ldots, Z_{ij}^{(k)}) \) be the \( k \) dimensional partial assignment vector of the first \( k \) treatments to unit \( i,j, 1 \leq k \leq K \). The units are assumed to be assigned treatments independently — \( Z_{ijK} \) is independent of \( Z_{i'j'K} \) for two different units \( i,j \) and \( i',j' \) across pairs, but the different treatments to a unit need not be assigned independently — \( Z_{ij}^{(k)} \) need not be independent of \( Z_{ij}^{(k')} \) for any \( k' \). A father of an exposed child may have poor hygiene because he is accustomed to work in an environment where exposure to lead is high, or he may have
good hygiene. Since we make no assumption about the dependence structure of $Z_{ijK}$, either of the above associations is allowed in this model. Let $Z_k = (Z_{11k}, Z_{12k}, \ldots, Z_{I2k})$ be the $2kI$ vector of first $k$ treatment assignments on $2I$ units.

The outcome for unit $ij$ is $R_{ij} = r_{ij}(Z_{ijK})$, determined from a set of $2^K$ potential outcomes, $r_{ij}(z_K)$ where $z_K \in \{0, 1\}^K$ (Neyman, 1923; Rubin, 1974). Only a single element of this set is observed. If there is a causal effect, e.g. in §6.2, an effect of occupational exposure to lead, then the elaborate theory states that $r_{ij}(z_K) > r_{ij}(z'_K)$ for $z_K, z'_K \in \{0, 1\}^K$ whenever $z_K \succ z'_K$ ($\succ$ denotes the partial ordering induced by coordinatewise ordering).

6.3.2. Assignment of treatment $Z_K$

As mentioned above, it is assumed that $Z_{ijK}$ is independent of $Z_{i'j'K}$ and that there is no interference between the units. This section defines the distribution of treatment exposure $Z_{ijK}$. The treatment assignment model is determined by the observed pre-treatment variables and the unmeasured confounders. This section also introduces the sensitivity parameters of our analysis.

Let $u_{ij1}, \ldots, u_{ijK}$ be $K$ unmeasured variables, $0 \leq u_{ijk} \leq 1$, $1 \leq k \leq K$ (Rosenbaum, 2002). Set $F = \{(r_{ij}(z_K), z_K \in \{0, 1\}^K), x_{ij}, u_{ij1}, \ldots, u_{ijK}; i = 1, \ldots, K, j = 1, 2\}$. We specify the distribution of $Z_{ijK}$ as the product of conditional distributions, i.e. $\Pr(Z_{ijK} = z_{ijK} \mid F) = \Pr(Z_{ij(1)} = z_{ij(1)} \mid F) \prod_{k \geq 2} \Pr(Z_{ij(k)} = z_{ij(k)} \mid F, Z_{ij(k-1)} = z_{ij(k-1)})$.

For the first treatment, treatment 1, we consider the model

$$\Pr(Z_{ij(1)} = 1 \mid F) = \frac{\exp(\theta_1(x_{ij}) + \gamma_1 u_{ij1})}{1 + \exp(\theta_1(x_{ij}) + \gamma_1 u_{ij1})}. \tag{6.1}$$

Here, $\theta_1()$ is an arbitrary unknown function and $\gamma_1 \geq 0$ is a sensitivity parameter, also unknown. Under this model, as units are matched so that $Z_{i1(1)} + Z_{i2(1)} = 1$, we have

$$\Pr(Z_{i1(1)} = 1 \mid F, Z_{i1(1)} + Z_{i2(1)} = 1) = \frac{\exp(\gamma_1 u_{ij1})}{\exp(\gamma_1 u_{ij2}) + \exp(\gamma_1 u_{ij1})}. \tag{6.2}$$
With $\Gamma_1 = \exp(\gamma_1)$, the odds ratio of treatment 1 satisfies $\Gamma_1^{-1} \leq \Pr(Z_{1i}^{(1)} = 1 \mid F, Z_{1i}^{(1)} + Z_{i2}^{(1)} = 1) \Pr(Z_{1i}^{(1)} = 0 \mid F, Z_{1i}^{(1)} + Z_{i2}^{(1)} = 1) \{\Pr(Z_{1i}^{(1)} = 0 \mid F, Z_{1i}^{(1)} + Z_{i2}^{(1)} = 1) \Pr(Z_{1i}^{(1)} = 1 \mid F, Z_{1i}^{(1)} + Z_{i2}^{(1)} = 1)\}^{-1} \leq \Gamma_1$. When $\Gamma_1 = 1$ ($\gamma_1 = 0$) the odds ratio is 1 and the probability of unit $ij$ getting treatment 1 in pair $i$ is a coin flip. Thus, $\Gamma_1$ is a parameter that measures the deviation from the random assignment of treatment 1 in the pairs.

Consider the model for $Z_{ij}^{(k)}$ as

$$\Pr(Z_{ij}^{(k)} = 1 \mid F, Z_{ij(k-1)} = z_{ij(k-1)}) = \frac{\exp(\theta_k(z_{ij(k-1)} + \gamma_k u_{ijk}))}{1 + \exp(\theta_k(z_{ij(k-1)} + \gamma_k u_{ijk})}, \quad (6.3)$$

for $k \geq 2$. As before, $\theta_k()$ is an unknown function and $\gamma_k \geq 0$ is a sensitivity parameter.

Upon conditioning on $Z_{k-1}$ the interpretation of $\gamma_k$ becomes clearer when we consider the distribution of $(Z_{1i}^{(k)}, Z_{i2}^{(k)}, \ldots, Z_{i2}^{(k)})$. Let $a_{k-1} \in \{0, 1\}^{k-1}$, consider the set of all units with $Z_{ij(k-1)} = a_{k-1}$; write it as $I_{k-1}(a_{k-1})$. Further write $|I_{k-1}(a_{k-1})| = n_{a_{k-1}}$ for the number of these units. Denote by $Z^{(k)}(I_{k-1}(a_{k-1}))$ the vector of length $n_{a_{k-1}}$ of $k$th treatment of the units in $I_{k-1}(a_{k-1})$ and by $u_k(I_{k-1}(a_{k-1}))$ the corresponding vector of $k$th unmeasured confounders, $u_{ijk}$’s. For $1 \leq m \leq n_{a_{k-1}}$, let $Z_{n_{a_{k-1}},m}$ be the binary vectors of length $n_{a_{k-1}}$ with $m$ ones and $n_{a_{k-1}} - m$ zeros. Then (6.3) implies

$$\Pr(Z^{(k)}(I_{k-1}(a_{k-1})) = z \mid F, Z_{k-1}, \sum_{ij \in I_{k-1}(a_{k-1})} Z_{ij}^{(k)} = m) = \frac{\exp(\gamma_k z^\top u_k(I_{k-1}(a_{k-1})))}{\sum_{\zeta \in Z_{n_{a_{k-1}},m}} \exp(\gamma_k \zeta^\top u_k(I_{k-1}(a_{k-1})))}, \quad \text{for } z \in Z_{n_{a_{k-1}},m}. \quad (6.4)$$

Irrespective of the value of $u_{ijk}$’s, if $\gamma_k = 0$ ($\Gamma_k := \exp(\gamma_k) = 1$), this probability is $(n_{a_{k-1}})^{-1}$, which indicates a randomized assignment of $m$ units to be treated with treatment $k$ among the units in $I(a_{k-1})$. The larger the value of $\Gamma_k$ is the bias in treatment $k$ is further from this random assignment.
6.3.3. $K$ tests for the causal hypothesis and their sensitivity to unmeasured confounding

The causal hypothesis has broad implications. When it is true, an exposure to the treatment, at any level, increases the outcome. This section reviews various nonparametric test statistics for the implications of the causal hypothesis and, using the treatment assignment model discussed in §6.3.2, also reviews the methods to assess the sensitivity of these tests to unmeasured confounders. Consider ranking of the responses by a preferred choice of ranking/scoring method for the $K$ tests. Let $q_{ijk}$ be the nonnegative score of unit $ij$ for test $k$, $k = 1, \ldots, K$. The scores are determined from the observed outcomes $(R_{11}, R_{12}, \ldots, R_{I2})$.

Fix $a = (a_1, \ldots, a_K) \in \{0, 1\}^K$ and let $a_{k-1} = (a_1, \ldots, a_{k-1})$, $2 \leq k \leq K$. For convenience we further write for $k = 1$, $k - 1 = 0$, $a_{k-1} = a_0 = \emptyset$. As in our discussion of §6.3.2, let $\mathcal{I}_{k-1}(a_{k-1})$ be the set of units with $Z_{ij(k-1)} = a_{k-1}$. Set $\mathcal{I}_0(a_0) = \mathcal{I}_0(\emptyset)$ to be the set of all $2I$ study units. Then we consider the following form the test statistics for the paired comparison on treatment 1

$$T_{1,a_0} = \sum_{i=1}^{I} \text{sgn}\left\{(Z_{ij(1)} - Z_{ij(2)})(R_{i1} - R_{i2})\right\}(q_{i11} + q_{i21}).$$

The function $\text{sgn}(x)$ is $-1$, $0$ or $1$ depending on $x < 0$, $x = 0$ or $x > 0$. Our test statistics for the effect of treatment $k \geq 2$ is

$$T_{k,a_{k-1}} = \sum_{ij \in \mathcal{I}_{k-1}(a_{k-1})} Z_{ij}^{(k)} q_{ijk}.$$

When $k = 1$ the test statistics is a pairwise comparison. In particular, if $q_{ij1} = q_{ij2}$ is the rank of absolute difference $|R_{i1} - R_{i2}|$ in the sorted list of the pairwise absolute differences, then $T_{1,a_0}$ is twice the Wilcoxon signed rank test statistics. When $k \geq 2$ the test is across pairs. But since it conditions on $a_{k-1}$, thus in particular fixes $Z_{ij}^{(1)}$ of all the units in $\mathcal{I}_{k-1}(a_{k-1})$, at most one unit from each pair is considered. Technically though there is no harm in scoring $ij \in \mathcal{I}_{k-1}(a_{k-1})$ as $q_{ijk}$ by also using outcomes of units $i'j' \notin \mathcal{I}_{k-1}(a_{k-1})$. 

117
Let $P_{k,a_{k-1}}$ be the p-value assessing the extent to which the test statistics $T_{k,a_{k-1}}$ provides evidence for an effect of treatment $k$. The null hypothesis, $H_0$, is Fisher’s sharp null so that $r_{ij}(z_{ijK}) = r_{ij}(z_{ijK}')$ for all $ij$ and $z_{ijK}, z_{ijK}' \in \{0,1\}^K$. If $T_{k,a_{k-1}}^{obs}$ is the observed value of the test statistics in the data then

$$P_{k,a_{k-1}} = \Pr(T_{k,a_{k-1}} \geq T_{k,a_{k-1}}^{obs} | F, Z_{k-1}, \sum_{ij \in I_{k-1}(a_{k-1})} Z_{ij}^{(k)}, H_0). \tag{6.5}$$

The test for the effect of exposure to $k$th treatment conditions on $Z_{k-1}$ and $\sum_{ij \in I_{k-1}(a_{k-1})} Z_{ij}^{(k)}$ as they are irrelevant for the effect (Kalbfleish, 1975; Helland, 1995). Conditioning on $H_0$ does not affect the treatment assignment distributions (6.1)–(6.4). If we could know $u_{ijk}$, we would calculate these p-values from the first principle using the probability distribution (6.2) if $k = 1$ and (6.4) if $k \geq 2$. The same is true if $\gamma_k = 0$. In the former of these two cases there is potentially bias from confounding variable, but these variables are known. In the second scenario there is no bias from unmeasured confounding and we use the conditional randomization distribution of the treatment $k$ for calculating the p-values.

However, the unmeasured confounders, $u_{ijk}$’s are just that — unmeasured. Thus, $P_{k,a_{k-1}}$ cannot be calculated if $\gamma_k > 0$. We calculate the maximum value of the p-value $P_{k,a_{k-1}}$ after fixing $\Gamma_k = \exp(\gamma_k)$, over the range of $u_{ijk}$; call this maximum $\overline{P}_{k,a_{k-1},\Gamma_k}$. The calculation is different between $\overline{P}_{1,a_0,\Gamma_1}$, the paired comparison for treatment 1, and $\overline{P}_{k,a_{k-1},\Gamma_k}$ for $k \geq 2$, between pair comparisons. Consider the paired comparison. Then

$$\overline{P}_{1,a_0,\Gamma_1} = \Pr(\sum_{i=1}^{I} s_i(q_{i11} + q_{i21}) \geq T_{1,a_0}^{obs} | F),$$

where $s_i$’s are independently distributed taking values 1 with probability $\Gamma_1/(1 + \Gamma_1)$ and $-1$ with probability $(1 + \Gamma_1)^{-1}$ if $R_{i1} \neq R_{i2}$ and $s_i \equiv 0$ if $R_{i1} = R_{i2}$ (Rosenbaum, 1987; 2002, §4.3).

The finite sample calculation of $\overline{P}_{k,a_{k-1},\Gamma_k}$, $k \geq 2$, is cumbersome. Recall $I_{k-1}(a_{k-1})$ is the
set of units with \( Z_{ij(k-1)} = a_{k-1} \). Let \( n_{a_{k-1}} = |\mathcal{I}_{k-1}(a_{k-1})| \) and \( m = \sum_{ij \in \mathcal{I}_{k-1}(a_{k-1})} Z_{ij}^{(k)} \).

Temporarily denote the units in \( \mathcal{I}_{k-1}(a_{k-1}) \) by \( \tilde{1}, \ldots, \tilde{m}_{a_{k-1}} \) so that the corresponding \( k \) scores are sorted in increasing order, \( q_{1k} \leq \cdots \leq q_{m_{a_{k-1}} k} \). There are \( 2^{m_{a_{k-1}}} \) values of \( u(\mathcal{I}_{k-1}(a_{k-1})) \) to maximize over. This can immediately be reduced to maximizing over only \( n_{i,a_{k-1}} - 1 \) of them. These \( u(\mathcal{I}_{k-1}(a_{k-1})) \)'s correspond to an \( l = 1, \ldots, n_{i,a_{k-1}} - 1 \), so that \( u_{1k} = \cdots = u_{lk} = 0 \) and \( u_{i(l+1)k} = \cdots = u_{m_{a_{k-1}} k} = 1 \). Still, the exact evaluation of the probabilities for these \( l \) instances is less than efficient. We consider the large sample approximation bound. It requires the following function

\[
C_k(a, b, c) = \sum_{l=\max(0,b+c-a)}^{\min(b,c)} \binom{c}{l} \binom{a-c}{b-l} e^{\gamma kl} \cdot 1(a \geq b, b > 0, c > 0).
\]

This function was discussed in Rosenbaum and Krieger (1990), equation (8). Let \( \Sigma_{l,a_{k-1}} \) be a symmetric matrix of size \( n_{a_{k-1}} \) defined as follows. The diagonal element of this matrix is \( \Sigma_{l,a_{k-1}}(\tilde{j}, \tilde{j}) = C_k(\mathcal{I}_{k-1}(a_{k-1}) - 1, m - 1, l) \{C_k(\mathcal{I}_{k-1}(a_{k-1}), m, l)\}^{-1} \) if \( \tilde{j} \leq l \) and \( \Sigma_{l,a_{k-1}}(\tilde{j}, \tilde{j}) = \Gamma_k C_k(\mathcal{I}_{k-1}(a_{k-1}) - 1, m - 1, l - 1) \{C_k(\mathcal{I}_{k-1}(a_{k-1}), m, l)\}^{-1} \) if \( \tilde{j} \geq l + 1 \). The \( (\tilde{j}, \tilde{j}') \)th off-diagonal element of this symmetric matrix is \( C_k(\mathcal{I}_{k-1}(a_{k-1}) - 2, m - 2, l) \{C_k(\mathcal{I}_{k-1}(a_{k-1}), m, l)\}^{-1} \) if \( \tilde{j} \leq l \) and \( \tilde{j}' \leq l \); it is \( \Gamma_k C_k(\mathcal{I}_{k-1}(a_{k-1}) - 2, m - 2, l - 1) \{C_k(\mathcal{I}_{k-1}(a_{k-1}), m, l)\}^{-1} \) if \( \tilde{j} \leq l \) and \( \tilde{j}' \geq l + 1 \); and it is \( \Gamma^2_k C_k(\mathcal{I}_{k-1}(a_{k-1}) - 2, m - 2, l - 2) \{C_k(\mathcal{I}_{k-1}(a_{k-1}), m, l)\}^{-1} \) if \( \tilde{j} \geq l + 1 \) and \( \tilde{j}' \geq l + 1 \). Then the mean of the test statistics for the unmeasured confounder \( l \) is

\[
\mu_{l,a_{k-1}} = \sum_{j=1}^{l} \frac{C_k(\mathcal{I}_{k-1}(a_{k-1}) - 1, m - 1, l) C_k(\mathcal{I}_{k-1}(a_{k-1}), m, l)}{C_k(\mathcal{I}_{k-1}(a_{k-1}), m, l)} q_{ij} \mu_{k-1} + \sum_{j=l+1}^{n_{a_{k-1}}} \frac{C_k(\mathcal{I}_{k-1}(a_{k-1}) - 1, m - 1, l - 1) C_k(\mathcal{I}_{k-1}(a_{k-1}), m, l)}{C_k(\mathcal{I}_{k-1}(a_{k-1}), m, l)} q_{ij} \mu_{k-1} - \mu_{l,a_{k-1}}.
\]

and the variance is

\[
\nu_{l,a_{k-1}}^2 = \sum_{\tilde{j}, \tilde{j}'=1}^{n_{a_{k-1}}} \Sigma_{l,a_{k-1}}(\tilde{j}, \tilde{j}') q_{ij} q_{ij'} - \mu_{l,a_{k-1}}^2.
\]

Then the asymptotically correct value, as \( I \to \infty \), of the maximum p-value for the \( k \)th test
statistics is (Rosenbaum, 2002, §4.6, §4.7)

\[ P_{k,a_{k-1},\Gamma_k} = 1 - \min_{l=1,\ldots,a_{k-1} - 1} \Phi^{-1}(\frac{T_{k,a_{k-1}} - \mu_{l,a_{k-1}}}{\nu_{l,a_{k-1}}}). \]

For various methods of sensitivity analysis in observational studies, see Cornfield et al. (1959), Egleston et al. (2009), Fogarty ans Small (2016), Fogarty and Hasegawa (2018), Gilbert et al. (2010), Hosman et al. (2010), Liu et al. (2013), and Yu and Gastwirth (2005). In particular see Rosenbaum (2018) for a comprehensive discussion and faster computation of \( P_{k,a_{k-1},\Gamma_k} \).

6.4. Evidence factors and pooling evidence

In the lead absorption study of §6.2 there are \( K = 5 \) tests with \( a = (1,1,1,1,1) \) or \( a = (1,1,1,1,0) \); the last coordinate is irrelevant for the design of the tests. The \( K \) test statistics are \( T_{k,a_{k-1}} \), \( 1 \leq k \leq K \). The previous section showed the computation of the maximum p-values for the these test statistics when the bias from unmeasured confounders is at most \( \Gamma_k \). These maximum p-values are denoted by \( P_{k,a_{k-1},\Gamma_k} \). Considered separately, for significance level \( \alpha \), the test using \( T_{k,a_{k-1}} \) is sensitive at level \( \Gamma_k \) if \( P_{k,a_{k-1},\Gamma_k} \geq \alpha \). This section establishes that these tests form evidence factors — they are biased by separate confoundings and they are nearly independent when the null is true.

**Proposition 3** Fix \( a \in \{0,1\}^K \). Under \( H_0 \), when the treatment assignment model is as (6.1) and (6.3), i.e. the bias in treatment \( k \) is at most \( \Gamma_k = \exp(\gamma_k) \)

\[ \Pr(\ P_{k,a_{k-1},\Gamma_k} \leq \alpha_k \forall k \geq 1 \ | \mathcal{F} ) \leq \prod_{k=1}^{K} \alpha_k. \]
Proof. We first note that \( \overline{P}_{k, a_{k-1}, r_k} \), which is the maximum value of \( P_{k, a_{k-1}, r_k} \) in (6.5) under model, is a function of \( F, Z_{k-1} \) and \( \sum_{ij \in I_{k-1}(a_{k-1})} Z_{ij}^{(k)} \). We write

\[
\Pr(\overline{P}_{k, a_{k-1}, r_k} \leq \alpha_k \forall k \geq 1 | F) = \Pr(\overline{P}_{1, a_0, r_1} \leq \alpha_1 | F) \times \prod_{k=2}^{K} \Pr(\overline{P}_{k, a_{k-1}, r_k} \leq \alpha_k | \overline{P}_{k', a_{k'-1}, r_{k'}} \leq \alpha_{k'} \forall k' \leq k - 1, F).
\]

Under \( H_0 \) and (6.1), \( \Pr(\overline{P}_{1, a_0, r_1} \leq \alpha_1 | F) \leq \alpha_1 \). Further for any \( k \geq 2 \), \( \Pr(\overline{P}_{k, a_{k-1}, r_k} \leq \alpha_k | \overline{P}_{k', a_{k'-1}, r_{k'}} \leq \alpha_{k'} \forall k' \leq k - 1, F) = E[E(1(\overline{P}_{k, a_{k-1}, r_k} \leq \alpha_k) | \overline{P}_{k', a_{k'-1}, r_{k'}} \leq \alpha_{k'} \forall k' \leq k - 1, F, Z_{k-1}, \sum_{ij \in I_{k-1}(a_{k-1})} Z_{ij}^{(k)})]. \) The outer expectation marginalizes over \( Z_{k-1} \) and \( \sum_{ij \in I_{k-1}(a_{k-1})} Z_{ij}^{(k)} \). Under \( H_0 \) and (6.3), by (6.5), the inner expectation is at most \( \alpha_k \). Combining these facts gives the required result. ■

**Theorem 12** Fix \( a \in \{0, 1\}^K \). Let \( f : [0, 1]^K \to (-\infty, \infty) \) be a function which is non-decreasing in its coordinates, i.e. \( f(x_1, \ldots, x_k, \ldots, x_K) \geq f(x_1, \ldots, x'_k, \ldots, x_K) \) for any \( x'_k \geq x_k \). Suppose \( U_1, \ldots, U_K \) are \( K \) i.i.d. random variables uniformly distributed on \([0, 1]\).

Under \( H_0 \), when the treatment assignment model is as in (6.1) and (6.3), for \(-\infty \leq x \leq \infty\),

\[
\Pr(f(\overline{P}_{1, a_0, r_1}, \ldots, \overline{P}_{K, a_{K-1}, r_K}) \leq x | F) \leq \Pr(f(U_1, \ldots, U_K) \leq x).
\]

**Proof.** The proof of the theorem follows from Proposition 3, along with Theorem 6.B.4 and Theorem 6.B.16 of Shaked and Shanthikumar (2007). A more general statement, Theorem 13, is proved in the appendix. ■

Proposition 3 shows that the joint distribution of the \( K \) p-values is stochastically larger than the uniform distribution on \( K \) dimensional hyper-cube. Thus, the tests are nearly independent in the sense of Theorem 12. The consequence of Theorem 12 is that usual methods of combining independent p-values can be used to pool evidence and report a single number for the evidence against the null that there is no causal effect. In particular, one can use Fisher’s method (Fisher, 1932) of combining p-values to calculate
\( P_{K}^{Fisher} = \Pr(\chi_{2K}^2 > -2 \sum_{k=1}^{K} \log P_{k,a_k-1,\Gamma_k}) \). The dependence of \( P_{K}^{Fisher} \) on \( \Gamma_k \)'s is suppressed here for convenience of notation. Theorem 12 implies that for any \( \alpha \in [0,1] \), when the biases are at most \( \Gamma_k \), under \( H_0 \), \( \Pr(P_{K}^{Fisher} \leq \alpha) \leq \alpha \). There are many such methods. Becker (1994) is a convenient reference for such methods. Zaykin et al. (2002)'s method deserves special mention. Zaykin et al. proposed a variant of the Fisher’s method by combining independent p-values using a truncated product. The test statistics is a product of those p-values that are smaller than some truncation point, \( \kappa \). Hsu et al. (2013) show that the truncated product with \( \kappa = 0.20 \) or \( \kappa = 0.10 \) often has higher power than Fisher’s method when applied to p-value bounds from a sensitivity analysis. The intuition is: the individual maximum p-values are not uniform but rather stochastically larger than a uniform distribution on \( [0,1] \), thus conservative.

6.5. Evidence from a partial conjunction of the tests: A quantification of the extent of corroboration

The pooled evidence from all the \( K \) tests has the benefit of ease of interpretation, yet it only provide information on whether at least one of the \( K \) tests support the alternative hypothesis, not whether a larger fraction support the alternative hypothesis. This section considers evidence from partial conjunctions of the tests. Throughout this section we fix \( a \in \{0,1\}^K \).

Fix \( k, 1 \leq k \leq K \). The null hypothesis for the effect of treatment \( k \) is the hypothesis that treatment \( k \) does not change the potential outcome of the units. Written formally \( H_{0,k} : R_{ij}(z_{ijK}) = R_{ij}(z'_{ijK}) \) for \( z_{ijK}, z'_{ijK} \in \{0,1\}^K \) if \( z_{ijK}^{(k')} = z'_{ijK}^{(k')} \) for all \( k' \leq k - 1 \); the alternative, \( H_{1,k} \), states that treatment \( k \) increases the response. The test statistics \( T_{k,a_k-1} \) tests for this null hypothesis. The global null \( H_0 \) is equivalent to \( \cap_{k=1}^{K} H_{0,k} \). Indeed, in (6.5) we can replace \( H_0 \) by \( H_{0,k} \), all arguments of §6.3.3 and §6.4 remain unchanged. The pooled evidence as in §6.4 is evidence against intersection of \( K \) nulls \( H_{0,k} \)'s. A small value of the pooled evidence tells us that we have evidence for at least one of these (one sided) alternatives. Consequently, it preserves the familywise error rate.
The global null $H_0$ is still false if at least one of the hypotheses is false, or at least $k$ of them are false. Is there evidence that at least $k$ of the $K$ hypotheses are false? Write, for $1 \leq k \leq K$

$$H_0^{k|K} : \bigcup_{t=k-K+1}^K \cap_{l \in \{t_1, \ldots, t_l\}} 1 \leq t_1 < \cdots < t_l \leq K \text{ } H_{0,t},$$

for the hypothesis that at most $K-k$ of the $K$ nulls are false. If $H_0^{k|K}$ is false then at least $k$ hypotheses are false. Specifically, $H_0^{1|K} \equiv H_0$. The evidence against $H_0^{k|K}$, i.e. evidence that at least $k$ of the null hypotheses are false, is found by looking at the largest $K-k+1$ p-values. Recall the p-values bounds were denoted by $(P_{1,a_0,\Gamma_1}, \ldots, P_{K,a_{K-1},\Gamma_K})$.

Let $\Gamma = (\Gamma_1, \ldots, \Gamma_K)$. We denote by $P_{(1)a,\Gamma} \leq \cdots \leq P_{(K)a,\Gamma}$, those $K$ values in increasing order.

Consider a function $g_k : [0,1]^{K-k+1} \rightarrow [0,1]$. Then the evidence against $H_0^{k|K}$ has the form

$$P_{a,\Gamma}^{k|K} = g_k(P_{(1)a,\Gamma}, \ldots, P_{(K)a,\Gamma}).$$

(6.6)

Theorem 13 is a general statement of Proposition 3 and Theorem 12 for any subset of the tests. The proof of Theorem 13 is given in the appendix. This theorem will be required to study the p-values $P_{a,\Gamma}^{k|K}$'s.

**Theorem 13** Fix $a \in \{0,1\}^K$. Let $\mathcal{K} = \{k_1, \ldots, k_{|\mathcal{K}|}\} \subseteq \{1, \ldots, K\}$. Under $\cap_{k \in \mathcal{K}} H_{0,t}$, when treatment assignment model is (6.1) and (6.3), but only for $k \in \mathcal{K}$, then for any non-decreasing function $f_{\mathcal{K}} : [0,1]^{|\mathcal{K}|} \rightarrow (-\infty, \infty)$, for $|\mathcal{K}|$ i.i.d. uniform $[0,1]$ random variables $U_1, \ldots, U_{|\mathcal{K}|}$, and $-\infty < x < \infty$, 

$$\Pr(f_{\mathcal{K}}(P_{k_1,a_{k_1-1},\Gamma_{k_1}}, \ldots, P_{k_{|\mathcal{K}|},a_{k_{|\mathcal{K}|-1},\Gamma_{k_{|\mathcal{K}|}}}) \leq x \mid \mathcal{F}) \leq \Pr(f_{\mathcal{K}}(U_1, \ldots, U_{|\mathcal{K}|}) \leq x).$$

The $k$th test become sensitive for bias level $\Gamma_k$ when $P_{k,a_{k-1},\Gamma_k} \geq \alpha$. The test for the partial conjunction hypothesis, $H_0^{k|K}$, is sensitive at bias level $\Gamma = (\Gamma_1, \ldots, \Gamma_k)$ if the pooled p-value is more than $\alpha$, $P_{a,\Gamma}^{k|K} \geq \alpha$. Using Theorem 13 the following proposition establishes
that $P_{a,\Gamma}^{k|K}$ in (6.6) is a p-value for testing $H_0^{k|K}$. Proposition 4 is equivalent to Theorem 1 of Benjamini and Heller (2008). See also Wang and Owen (2017) for related results.

**Proposition 4** Fix $a \in \{0, 1\}^K$. Consider model (6.1) and (6.3). Let $g_k : [0, 1]^{K-k+1} \rightarrow [0, 1]$ be a coordinatewise nondecreasing function in (6.6). Suppose, $\Pr(g_k(U_k, \ldots, U_K) \leq \alpha)$ for some $\alpha \in [0, 1]$, where $U_1, \ldots, U_K$ are i.i.d uniform random variables on $[0, 1]$. Then, under $H_0^{k|K}$

$$\Pr(P_{a,\Gamma}^{k|K} \leq \alpha \mid \mathcal{F}) \leq \alpha. \quad (6.7)$$

**Proof.** Recall, $H_0^{k|K} : \bigcup_{t \in K-K-k+1} \bigcap_{t \in \{t_1, \ldots, t_l\}, 1 \leq t_1 < \cdots < t_l \leq K} H_{0, t}$. Fix, $1 \leq t_1 < \cdots < t_l \leq K$ for some $l \geq K-k+1$ and set $\mathcal{K} = \{t_1, \ldots, t_{K-k+1}\}$. Then, $\bigcap_{t \in \mathcal{K}} H_{0, t}$ implies $\cap_{t \in \mathcal{K}} H_{0, t}$. By (6.6), with the fact that $g_k$ is coordinatewise nondecreasing and Theorem 13, respectively, we bound the probability in (6.7) by

$$\Pr(g_k(P_{t_1, a_{t_1-1, t_1}, \ldots, P_{t_{K-k+1}, a_{t_{K-k+1}-1, t_{K-k+1}}}) \leq \alpha \mid \mathcal{F}) \leq \Pr(g_k(U_1, \ldots, U_{K-k+1}) \leq \alpha) \leq \alpha.$$

- \]

**Proposition 5** Consider $K$ functions, $g_k : [0, 1]^{K-k+1} \rightarrow [0, 1]$, $1 \leq k \leq K$. Assume the following, for i.i.d. uniform $[0, 1]$ random variables $U_1, \ldots, U_K$, for all $k = 1, \ldots, K$

(a) $g_k$ is nondecreasing in its coordinates.

(b) $\Pr(g_k(U_k, \ldots, U_K) \leq \alpha)$ for some $\alpha \in [0, 1]$.

(c) $g_k(x_k, x_{k+1}, \ldots, x_K) \leq g_{k+1}(x_{k+1}, \ldots, x_K)$ for all $x_{k+1}, \ldots, x_K \in [0, 1]$ and $x_k \leq \min\{x_{k+1}, \ldots, x_K\}$.

Condition (c) is void if $k = K$. Fix $a \in \{0, 1\}^K$. Suppose we reject $H_0^{k|K}$ if $P_{a,\Gamma}^{k|K} = g_k(P_{(k)a, \Gamma}, \ldots, P_{(K)a, \Gamma})$ is less than $\alpha$. Under model (6.1) and (6.3), the probability of
rejecting any true null hypothesis among \( \{ H_0^{k|K}; k = 1, \ldots, K \} \) is at most \( \alpha \).

**Proof.** Since, \( H_0^{k|K} \) is the hypothesis that at most \( k - 1 \) nulls are false, they satisfy \( H_0^{1|K} \subseteq \cdots \subseteq H_0^{K|K} \). Further, condition (c) implies \( P_a^1 \leq \cdots \leq P_a^K \). This is because, by (c), for \( k = 1, \ldots, K - 1 \), \( P_a^k = g_k(\mathcal{P}(k)_{a, \Gamma}, \ldots, \mathcal{P}(K)_{a, \Gamma}) \leq g_{k+1}(\mathcal{P}(k+1)_{a, \Gamma}, \ldots, \mathcal{P}(K)_{a, \Gamma}) = P_a^{k+1} \).

If there is no true null among \( \{ H_0^{k|K}; k = 1, \ldots, K \} \) there is nothing to prove. Otherwise, let \( k \) be the smallest number such that \( H_0^{k|K} \) is true. Consequently, \( H_0^{1|K}, \ldots, H_0^{k-1|K} \) are false. Then a false rejection implies rejection of a null hypothesis \( H_0^{k|K} \) which is true and \( k' \geq k \) with \( P_a^{k|K} < \alpha \). From the ordering of the p-values noted above, it implies \( P_a^{k|K} < \alpha \).

Hence the probability of rejecting any true null hypothesis among \( \{ H_0^{k|K}; k = 1, \ldots, K \} \) is bounded by \( \Pr(P_a^{k|K} < \alpha \mid \mathcal{F}, H_0^{k|K}) \). This is at most \( \alpha \) by condition (a) and (b) using Proposition 4.  

Condition (c) of Proposition 5 is satisfied by Simes’ method of combining p-values (Simes, 1986). To see this, consider \( 0 \leq x_k \leq x_{k+1} \leq \cdots \leq x_K \leq 1 \). Simes’ method uses the function \( g_k(x_k, x_{k+1}, \ldots, x_K) = \min_{l=1,\ldots,K-k+1} l^{-1}(K-k+1)x_{k+l-1} \) in calculating \( P_a^{k|K} \) using (6.6). Accordingly, \( g_{k+1}(x_{k+1}, \ldots, x_K) = \min_{l=1,\ldots,K-k} l^{-1}(K-k)x_{k+l} = \min_{l=2,\ldots,K-k+1} (l-1)^{-1}(K-k)x_{k+l-1} \). It follows that

\[
g_k(x_k, x_{k+1}, \ldots, x_K) = \min_{l=1,\ldots,K-k+1} l^{-1}(K-k+1)x_{k+l-1} \\
\leq \min_{l=2,\ldots,K-k+1} l^{-1}(K-k+1)x_{k+l-1} \\
= \min_{l=2,\ldots,K-k+1} \{(l-1)l^{-1}(K-k+1)(K-k)^{-1}(l-1)^{-1}(K-k)x_{k+l-1} \} \\
\leq \min_{l=2,\ldots,K-k+1} (l-1)^{-1}(K-k)x_{k+l-1} \\
= g_{k+1}(x_{k+1}, \ldots, x_K).
\]

Although, this condition may not be satisfied generally by any method of combining p-values. For example, it is not satisfied by Fisher’s method. To see this let \( K = 2 \), \( x_1 = x_2 = 0.5 \). Then \( g_1(x_1, x_2) = \Pr(\chi_4^2 > -2 \log x_1 \cdot x_2) \approx 0.596 > 0.5 = \Pr(\chi_2^2 > -2 \log x_2) = \)

\[
0.596 > 0.5
\]

125
$g_2(x_2)$. The following proposition lists other methods that satisfies the conditions (a)–(c) of Proposition 5. The first one in this list looks only at the minimum p-value $P_k$ for testing $H_0^{K/k}$. This ‘minimum p-value’ method is fairly well known in the statistics literature. The following method is Stouffer’s method which is popular in the meta-analysis literature (Stouffer et al., 1949). The last method in this list is a modification of ‘additive p-value method’ of Edgington (1972).

**Proposition 6** Conditions (a)–(c) of Proposition 5 are satisfied by each of the following specifications of $g_k$.

1. *(minimum p-value method)* $g_k(x_k, \ldots, x_K) = 1 - (1 - \min\{x_k, \ldots, x_K\})^{K-k+1}$.

2. *(sum of z’s)* $g_k(x_k, \ldots, x_K) = 1 - \Phi\left(\frac{\Phi^{-1}(1-x_k) + \cdots + \Phi^{-1}(1-x_K)}{\sqrt{K-k+1}}\right)$.

3. *(modified additive p-value method)* With $A_k = x_k + \cdots + x_K$, $g_k(x_k, \ldots, x_K) = \left(\min\left\{\frac{A_k^{K-k+1}}{(K-k+1)!}\right\}^{1(A_k \leq c_k)}\right)$ where $c_k = (K - k + 1)(1 - (K - k + 2)^{-1})^{K-k+1}$.

The proof of this proposition is given in the appendix. In might often be useful to weight the p-values when combining them. However, the validity of the combined p-value for the partial conjunction hypothesis would usually require the weights to be predetermined. Also, the optimal choice of the weights could depend on the specific problem (Chen, 2011; Lancaster, 1961; Liptak, 1958; Whitlock, 2005; Zaykin, 2011). We do not discuss the various methods of weighted combinations in this chapter.

The rest of this section considers the sensitivity analysis to unmeasured confounding over the multiple sensitivity parameters. There are $K$ sensitivity parameters, $\Gamma_1, \ldots, \Gamma_K$. We gradually establish that the proposed sensitivity analysis for testing of partial conjunction of the hypotheses will control for the familywise error rate.

In the sensitivity analysis one first fixes a range of values of the bias parameters. Let $1 = \Gamma_1 < \cdots < \Gamma_{1S_1}$ be the range of values for the bias parameter $\Gamma_1$ for bias in treatment 1; $1 = \Gamma_{k1} < \cdots < \Gamma_{kS_k}$ is the range of values for the bias parameter $\Gamma_k$ for treatment $k$. 

126
Let $\mathcal{J} = \{ \Gamma = (\Gamma_{1s_1}, \ldots, \Gamma_{Ks_K}) : 1 \leq s_1 \leq S_1; \ldots; 1 \leq s_K \leq S_K \}$. The goal is to find the least amount of bias that could explain an observed association. We denote by $H_{0,\Gamma}^{k|K}$ the conjunction of the hypothesis $H_{0}^{k|K}$ and that the bias is at most $\Gamma$. The statement that — the bias is at most $\gamma_k = \log \Gamma_k$ for some set of unmeasured confounders $u_{ijk}$’s. The following theorem says that the maximum error of the multi-parameter sensitivity analysis using $P_{a,\Gamma}$’s is bounded by $\alpha$.

**Theorem 14** Fix $k$, $1 \leq k \leq K$. Consider the set of sensitivity parameters $\mathcal{J} = \{ \Gamma = (\Gamma_{1s_1}, \ldots, \Gamma_{Ks_K}) : 1 \leq s_1 \leq S_1; \ldots; 1 \leq s_K \leq S_K \}$. Assume the conditions of Proposition 4. Fix $a \in \{0,1\}^K$. Consider the procedure that rejects $H_{0,\Gamma}^{k|K}$ for $\Gamma \in \mathcal{J}$ if $P_{a,\Gamma}^{k|K} < \alpha$. Then the probability of rejecting any true null hypothesis among the set of hypotheses $\{H_{0,\Gamma}^{k|K} ; \Gamma \in \mathcal{J} \}$ is at most $\alpha$.

**Proof.** Note first that $H_{0,\Gamma}^{k|K} \subseteq H_{0,\Gamma'}^{k|K}$ for $\Gamma' \succ \Gamma$. This is true since a bias of at most $\Gamma_k$ implies bias at most $\Gamma'_k$ for $\Gamma_k \leq \Gamma'_k$. Let $\overline{\Gamma} \in \mathcal{J}$ be such that $H_{0,\overline{\Gamma}}^{k|K}$ is true and if $\Gamma \in \mathcal{J}$ and $H_{0,\Gamma}^{k|K}$ is true then $\Gamma \succ \overline{\Gamma}$. $\overline{\Gamma}$ might be empty; in which case there is nothing to prove.

Next, we note that $P_{a,\Gamma}^{k|K}$ is increasing in $\Gamma$; $P_{a,\Gamma}^{k|K} \leq P_{a,\Gamma'}^{k|K}$ for $\Gamma \leq \Gamma'$. A rejection of a true null hypothesis when the corresponding maximum p-value is less than $\alpha$, implies $P_{a,\overline{\Gamma}}^{k|K} < \alpha$. Thus, the probability of rejecting any true null hypothesis is upper bounded by $\Pr(P_{a,\overline{\Gamma}}^{k|K} < \alpha)$, which is at most $\alpha$ by Proposition 4. ■

The following corollary to the theorem considers a sensitivity analysis with the same bias parameter for all the factors. The proof of the following two corollaries are given in the appendix.

**Corollary 2** Assume the same conditions as in Theorem 14, except let $\mathcal{J} = \{ \Gamma = \Gamma_1(1, \ldots, 1) : 1 = \Gamma_1 < \Gamma_2 < \cdots < \Gamma_L \}$. Fix $a \in \{0,1\}^K$ and $k$, $1 \leq k \leq K$. Consider the testing procedure that rejects $H_{0,\Gamma}^{k|K}$ for $\Gamma \in \mathcal{J}$ if $P_{a,\Gamma}^{k|K} < \alpha$. Then the probability of rejecting any true null hypothesis among the set of hypotheses $\{H_{0,\Gamma}^{k|K} ; \Gamma \in \mathcal{J} \}$ is at most $\alpha$.
The final corollary combines the situations of Proposition 5 and Theorem 14.

**Corollary 3** Assume that conditions (a)–(c) of Proposition 5 are satisfied and assume the structure of \( J \) either as in Theorem 14 or as in Corollary 2. Fix \( a \in \{0, 1\}^K \). Consider the procedure that rejects \( H_{0,\Gamma}^{k,K} \) for \( \Gamma \in J \) if \( P_{a,\Gamma}^{k,K} < \alpha \). Then the probability of rejecting any true null hypothesis among \( \{H_{0,\Gamma}^{k,K} : 1 \leq k \leq K, \Gamma \in J\} \) is at most \( \alpha \).

### 6.6. Comparison of combining methods

**6.6.1. Settings under which power of sensitivity analysis is judged**

In a sensitivity analysis to unmeasured confounding, there are some situations in which it is clear what we would like a procedure to do and some situations in which the desired answer is unclear. An example of one of the latter situations is when there is large bias from unmeasured confounding and a treatment effect — we are nearly assured to reject the null for moderate values of the sensitivity parameter, but, such a rejection decision is not unambiguously sought after as we would also have rejected the null with moderate bias when the null is indeed true. One of the former situations, in which we are clear about the desired answer of the sensitivity analysis, is when there is a treatment effect and no bias from unmeasured confounding. In this situation, a sensitivity analysis with a chosen value of the sensitivity parameter checks whether we are still able to reject the null, allowing for the level of bias given by the sensitivity parameter. It is desired then that a method is not fooled by moderate values of the sensitivity parameter and rejects the null. This situation has been called the “favorable situation” and is the situation under which power of sensitivity analysis has been evaluated (Rosenbaum, 2010; Hansen et al., 2014).

One might wonder why we evaluate the power of a sensitivity analysis under a setting in which there is actually no bias from unmeasured confounding when the sensitivity analysis is worried about bias. The reason is that, in most observational studies, we are worried about bias and cannot know that there is no bias, but we would like to have high power to say that we have evidence for a treatment effect that is insensitive to moderate bias if in
fact there is a treatment effect and no bias.

In §6.6.2, we will analyze the asymptotics of power of sensitivity analysis when the sample size goes to infinity. There we provide a characterization of the asymptotically optimal choice of combining method, and find asymptotically optimal combining methods. In §6.6.3, we compare the combining methods in their power of sensitivity analysis using a simulation study. Since in practice we only have a finite sample, looking at the power of sensitivity analysis for finite samples might give us more guidance about the choice of method for analysis.

6.6.2. Asymptotically optimal tests

When there is a treatment effect and no unmeasured confounding, a method is preferred that can withstand larger bias in sensitivity analysis. When the sample size goes to infinity, this threshold of the sensitivity parameter is quantified as the design sensitivity of the method (Rosenbaum, 2004; Rosenbaum, 2010; Hsu et al., 2013; Hansen et al., 2014; Zhao, 2018). However, for partial conjunction testing from $K$ evidence factors, design sensitivity for the various combining methods is a crude criterion of comparison. As we will see in Proposition 7 below, most combining methods have the same design sensitivity. Instead, we look at the rate of rejection for the combining methods in their sensitivity analysis when there is treatment effect and no unmeasured confounding. This rate of rejection is the Bahadur slope of a sensitivity analysis (Rosenbaum, 2015). The ratio of the slopes of two competing methods of analysis is called the Bahadur efficiency of sensitivity analysis. A method with larger slope needs a smaller sample size to make the desired decision with high probability (Bahadur, 1967; Rosenbaum, 2015; Ertefaie et al., 2018). In the following, we show that Fisher’s method and the truncated product method are optimal in this regard. Put differently, Fisher’s method (Fisher, 1932) and the truncated product method (Zaykin, 2002) have Bahadur efficiency of sensitivity analysis one, relative to each other, and have efficiency at least one, relative to any other combining method.
We first introduce some notation to facilitate the discussion. Recall, the partial conjunction p-values are defined for a set of functions \((g_1, \ldots, g_K)\) where \(g_k : [0, 1]^{K-k+1} \rightarrow [0, 1]\), \(k = 1, \ldots, K\), as
\[
P_{a,\Gamma}^{k|K} = g_k(\overline{P}_{1|a,\Gamma}, \ldots, \overline{P}_{K|a,\Gamma}).
\]
Here, \(\overline{P}_{1|a,\Gamma} \leq \cdots \leq \overline{P}_{K|a,\Gamma}\) are the ordered values of \(\overline{P}_{1, a_1, \Gamma_1}, \ldots, \overline{P}_{K, a_{K-1}, \Gamma_{K-1}}\). Now we emphasize the choice of the combining functions by denoting \(g = (g_1, \ldots, g_K)\) and using \(P_{a,\Gamma}^{k|K}(g)\) to denote the above quantity. We use the notation \(ef = (f_1, \ldots, f_K)\) to denote Fisher’s combining functions. That is, the \(k\)th function in \(ef\) is \(f_k(x_k, \ldots, x_K) = \Pr(\chi^2_{2(K-k+1)} > -2 \sum_{j=k}^K \log x_j)\). The optimality statement made in this section is an asymptotic statement. We must think of \(\overline{P}_{k, a_{k-1}, \Gamma_k}\) as function of \(I\), the number of pairs. Consequently, \(P_{a,\Gamma}^{k|K}(g)\) is also a function of \(I\). These dependencies will not be made explicit below. The asymptotic here is with \(K\) fixed and \(I\) going to infinity.

Consider the situation where there is an effect, i.e., some of the \(K\) hypotheses \(H_0^{(k)}\) are false. Suppose, there is no unmeasured confounding. We noted that the desired result of a sensitivity analysis, in this situation, is to be able to reject the null. Suppose \(H_0^{(k)}\) is false. The maximum p-value for the \(k\)th factor is \(\overline{P}_{k, a_{k-1}, \Gamma_k}\). For any sample size, as \(\Gamma_k \rightarrow \infty\) this maximum p-value \(\overline{P}_{k, a_{k-1}, \Gamma_k} \rightarrow 1\), a formal statement for the known fact that any treatment effect, however large, can be explained by large enough bias. The design sensitivity for this factor is the bias level \(\tilde{\Gamma}_k\) such that \(\overline{P}_{k, a_{k-1}, \Gamma_k} \rightarrow 0\) for \(\Gamma_k < \tilde{\Gamma}_k\) and \(\overline{P}_{k, a_{k-1}, \Gamma_k} \rightarrow 1\) for \(\Gamma_k > \tilde{\Gamma}_k\); the limit here is with \(I \rightarrow \infty\).

Now we look at the sensitivity analysis for the partial conjunctions of these evidence factors. The following proposition studies the design sensitivity of this multi-parameter sensitivity analysis, and concludes that most methods are indistinguishable in this regard.

**Proposition 7** Take any combining method \(g\). Suppose, \(g_k(0, \ldots) = 0\) and \(g_k(1, \ldots, 1) = 1\) and \(g_k\) is continuous at \([0, 1]^{K-k+1}\). With the sensitivity parameter \(\Gamma = (\Gamma_1, \ldots, \Gamma_K)\) for the partial conjunction testing, we have \(P_{a,\Gamma}^{k|K}(g) \rightarrow 1\) if \(\bar{\Gamma}_I < \Gamma_1\) for \(K - k + 1\) many \(\Gamma_I\).
Also, \( P_{a, \Gamma}^{k|K}(\mathbf{g}) \to 0 \) if \( \Gamma_l < \tilde{\Gamma}_l \) for at least \( k \) many \( \Gamma_l \) and \( \Gamma_l \neq \tilde{\Gamma}_l \) for all \( l \).

The following theorem says that, in the class of functions for \( \mathbf{g} \) considered in §6.5, Fisher’s method, \( \mathbf{ef} \), has the optimal Bahadur slope.

**Assumption:** A sequence of numbers \( c(I) \) satisfies \( c(I) \to \infty \) as \( I \to \infty \). As \( I \) increases to infinity, \( c(I)^{-1} \log P_{k, a_{k-1}, r_k} \to -r_k(\Gamma_k) \) almost surely, where \( r_k(\Gamma_k) \in [0, \infty] \), for \( k = 1, \ldots, K \). We call \( r_k(\Gamma_k) \) the slope of test \( k \) at \( \Gamma_k \).

**Theorem 15** Consider any set of \( K \) combining functions \( \mathbf{g} = (g_1, \ldots, g_K) \) such that each \( g_k \) is coordinatewise nondecreasing and satisfies \( \Pr(g_k(U_k, \ldots, U_K) \leq \alpha) \leq \alpha \), for any \( \alpha \in [0, 1] \), for i.i.d. uniform(0,1) random variables \( U_1, \ldots, U_K \); \( k = 1, \ldots, K \). We have, for Fisher’s combining method \( \mathbf{ef} = (f_1, \ldots, f_K) \),

\[
\lim_{I \to \infty} c(I)^{-1} \log p_{a, \Gamma}^{k|K}(\mathbf{ef}) \leq \lim_{I \to \infty} c(I)^{-1} \log p_{a, \Gamma}^{k|K}(\mathbf{g}) \quad \text{for } k' \leq k
\]

almost surely for \( k, k' = 1, \ldots, K \).

The assumption talks about the Bahadur slope of sensitivity analysis for the individual factors. Rosenbaum (2015) provides a detailed discussion on the existence and calculation of the limit. The limit depends on the choice of the test statistic, the joint distribution of the potential outcomes for the units, and the distribution of the treatment assignment.

The above assumption and the theorem while general also allow us to consolidate several important implications.

Following Proposition 7, our interest is in the case when we are able to reject the null in the sensitivity analysis, when in truth there is an effect. This is the case for a sensitivity parameter \( \Gamma \) with some of the bias levels less than the design sensitivity. Let \( \tilde{k} \) be the number of \( \Gamma_l \) with \( \Gamma_l < \tilde{\Gamma}_l \). Any method in Proposition 7 will reject \( H_0^{k|K} \) whenever \( k \leq \tilde{k} \), as the sample size goes to infinity. The rate of rejection is used in Theorem 15 to tell the combining methods apart. The following Proposition finds the slope of Fisher’s method.
This slope is the same as that of the truncated product method with a truncation level \( z \), and is at least as large as any other method that satisfies the conditions of Theorem 15.

**Proposition 8** Suppose there is no unmeasured confounding and \( H_0 \) is false. Let the design sensitivity of the test \( k \) be \( \tilde{\Gamma}_k \). Consider a sensitivity analysis with sensitivity parameter \( \Gamma \) such that \( \Gamma_k \neq \tilde{\Gamma}_k \) for all \( k \). Let \( \tilde{k} \) be the number of \( \Gamma_l \) with \( \Gamma_l < \tilde{\Gamma}_l \). Finally, let \( r_1(\Gamma) \leq \cdots \leq r_K(\Gamma) \) are ordered values of \( r(\Gamma_1), \ldots, r_K(\Gamma_K) \). We have 
\[
\lim_{I \to \infty} c(I)^{-1} \log P_{a,\Gamma}^{k|\Gamma}(\cdot|\cdot) = -1(k \leq \tilde{k}) \sum_{k-k+1}^{K} r_0(\Gamma) \Gamma. 
\] The truncated product method with \( z \in (0,1] \) has the same slope as Fisher’s method.

6.6.3. Simulation study: Finite sample power of sensitivity analysis

Section 6.5 discussed various choices of the function \( g_k \), which is used to define \( P_{a,\Gamma}^{k|\Gamma} \). In this section we compare these combining methods in their power of sensitivity analysis in finite samples using a simulation study.

In the simulation setting we set \( I = 150 \) and \( K = 5 \). Treatment \( k \) has an additive effect \( \beta_k \) and we assume a standard normal variate for the base response in the absence of any treatment. Thus, when a unit has been assigned treatment \((z_1, \ldots, z_K)\), a binary vector of length \( K \), the response of that unit is \( \sum_{k=1}^{K} z_k \beta_k + N(0,1) \). We simulate a treatment assignment which is random, thus within each pair, each unit has probability 1/2 of getting the first treatment. Further, the treatments are simulated to be independent of each other in a way that \( Z_{ij}^{(k)} \overset{i.i.d.}{\sim} Bernoulli(0.6) \) for \( 2 \leq k \leq 4 \) and \( Z_{ij}^{(5)} \overset{i.i.d.}{\sim} Bernoulli(0.5) \).

In the power of sensitivity analysis, we look at the simulated power of of rejecting \( H_0^{k|\Gamma} \) for the various methods when we assume various \( \Gamma \) values for bias. A method is less sensitive if, in the presence of a treatment effect, it maintains power to detect that treatment effect at higher values of \( \Gamma \) (Rosenbaum, 2004). We take \( a = (1,1,1,1,1) \) as in the §6.2. The basic tests use Wilcoxon’s paired sample and two sample statistics. These simulation results are presented in Table 12, where each sampling situation was replicated 15,000 times, so that a binomial proportion has a standard error less than \( \sqrt{0.25/15000} \approx 0.004 \).
The four methods compared in the simulation are Holm-Bonferroni method (henceforth Holm’s), Simes’ method, the modified additive p-value method (henceforth SumP) and the truncated product method. Holm’s method ignores the near independence of the separate analyses established in Theorem 13. For Holm’s method \( g_k(x_k, \ldots, x_K) = (K - k + 1)x_k \) (Holm, 1979). Simes’ method and the SumP method satisfy the desired conditions of Proposition 5, Holm’s method does not. For the truncated product method we consider the familiar level of truncation \( \alpha = 0.2 \). This method was further modified to redefine \( P^{k|K}_{a, \Gamma} = \max\{P^{1|K}_{a, \Gamma}, \ldots, P^{k|K}_{a, \Gamma}\} \) for \( k = 1, \ldots, K \), so that it provided monotone p-values, required in Proposition 5. In Table 12, a simulated power of 0 is replaced by a blank cell for ease of viewing.

There are at least two ways of reading Table 12. First, we look at each of the methods individually and compare the various scenarios of treatment effect. Note that, the power for each of the methods decrease as we read the table from right to left, increasing the value of \( k \), and top to bottom in each scenario, increasing the value of \( \Gamma \). The null case of no treatment effect, scenario 1, is a check that the analysis is performed at level of significance 0.05 and the methods control the type 1 error. Across the scenarios, moving from the null scenario to the scenario where each treatment has an effect of size 0.25 (scenario 3), the simulated power increases for each of the methods. The power of rejecting at least 3 basic hypotheses out of 5, \( H_0^{3|5} \), for \( \Gamma = 1 \), is 9\% for SumP method in Scenario 2 and 32\% in Scenario 3. The corresponding numbers are 5\% and 10\% for the Simes’ method, and 7\% and 23\% for the truncated product method.

Consider a second perspective to Table 12. We compare the methods within the various scenarios. The power of the SumP method is much smaller in rejecting \( H_0^{1|5} \) \((k = 1)\) compared to the other methods. The power, in scenario 2 with \( \Gamma = 1 \), is 57\% for SumP compared to 99\% for Holm’s, Simes’, and the truncated product method. Also, in terms of the maximum bias level of sensitivity analysis a method can tolerate, (which one can read by looking at how far down the numbers go before vanishing in each column) Holm’s and Simes’
Table 12: Simulation results for power of sensitivity analysis evaluated at level 0.05. Numbers are out of 100. A cell value is the percentage of times the decision that at least $k$ many $H_{0,8}$ are false is made, with $\Gamma_1 = \cdots = \Gamma_5 = \Gamma$, out of 15000 simulations. Empty cells represent the value 0. $tP = \text{truncated product method with truncation level } \varepsilon = 0.20$; $sP = \text{the modified additive p-value method in Proposition 6}; \text{Si} = \text{Simes' method}; \text{HB} = \text{Holm-Bonferroni method}.$

<table>
<thead>
<tr>
<th>$k \to$</th>
<th>$\Gamma$</th>
<th>5</th>
<th>4</th>
<th>3</th>
<th>2</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>tP</td>
<td>sP</td>
<td>Si</td>
<td>HB</td>
<td>tP</td>
<td>sP</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Scenario 1: (null case) $\beta_1 = \cdots = \beta_5 = 0$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.2</td>
<td>1</td>
<td>2</td>
<td>7</td>
<td>9</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>1.4</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>1.6</td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.8</td>
<td></td>
<td></td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Scenario 2: $\beta_1 = \beta_2 = \beta_3 = 0.25, \beta_4 = \beta_5 = 0$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>6</td>
<td>10</td>
<td>2</td>
<td>1</td>
<td>23</td>
<td>32</td>
</tr>
<tr>
<td>1.2</td>
<td>1</td>
<td>3</td>
<td>8</td>
<td>15</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>1.4</td>
<td>1</td>
<td>3</td>
<td>6</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>1.6</td>
<td>1</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.8</td>
<td></td>
<td></td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>2.2</td>
<td></td>
<td></td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2.4</td>
<td></td>
<td></td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.6</td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.8</td>
<td></td>
<td></td>
<td>5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3.6</td>
<td></td>
<td></td>
<td>3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

method are less sensitive when $k = 1$ for both scenario 2 and 3. The story is somewhat reversed for larger $k$. For example, consider $k = 3$ or $H_0^{35}$ in scenario 3. The simulated power for $\Gamma = 1$ is highest for SumP (32%) and lowest for Holm’s method (8%) and second lowest for Simes’ method (10%); for the truncated product it is 23%. Further, SumP is less sensitive (sensitive at $\Gamma = 2.2$) compared to Simes’, and Holm’s method (sensitive at $\Gamma = 1.8$).
Table 13: The p-values, under the assumption of no unmeasured confounding, for testing the hypothesis that at least \( k \) many \( H_0 \)'s are false in the lead absorption study. \( K = 5 \) and \( \Gamma_1 = \cdots = \Gamma_5 = 1 \). SumP = the modified additive p-value method in Proposition 6.

<table>
<thead>
<tr>
<th>( k )</th>
<th>Simes’ SumP</th>
<th>Fisher’s ((\alpha = 0.20))</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>0.420036</td>
<td>0.420036 1</td>
</tr>
<tr>
<td>4</td>
<td>0.191846</td>
<td>0.133107 0.169691 0.193477</td>
</tr>
<tr>
<td>3</td>
<td>0.028322</td>
<td>0.024172 0.015168 0.017172</td>
</tr>
<tr>
<td>2</td>
<td>0.015242</td>
<td>0.003268 0.000739 0.000795</td>
</tr>
<tr>
<td>1</td>
<td>0.000348</td>
<td>0.000346 1.41 ( \cdot ) ( 10^{-6} ) 1.57 ( \cdot ) ( 10^{-6} )</td>
</tr>
</tbody>
</table>

To summarize, no one method is victorious. But it seems Simes’ or Holm’s method is a poor choice as they lose their power fast going from right to left of the table. Holm’s method essentially looks at the individual p-values and does not pool them, thus it often misses that there is evidence for some fraction of the nulls not being true when each test does not have sufficient power. While SumP has a much smaller power in providing evidence that at least one of the nulls is false, it retains a lot of its power when looking for more pieces of evidence (going right to left). The truncated product method seems be a fair compromise based on these simulations.

6.7. Revisiting the lead absorption study

The p-values for the five tests were reported in §6.2 for the causal hypothesis that occupational exposure to lead increases the lead level in the blood of the children. If there is no bias due to unmeasured confounding, i.e., assuming \( \Gamma_1 = \cdots = \Gamma_5 = 1 \), these p-values are \( P_1 = 2.69 \cdot 10^{-5} \), \( P_2 = 3.81 \cdot 10^{-3} \), \( P_3 = 9.59 \cdot 10^{-2} \), \( P_4 = 9.44 \cdot 10^{-3} \), and \( P_5 = 0.42 \). The p-values for the tests for partial conjunction of the hypotheses are given in Table 13. This table reports the results from four methods of pooling evidence. Qualitatively, the results from the four methods are similar. At \( \alpha = 0.05 \), we have evidence for rejecting at least 3 out of 5 basic nulls. The p-values from Fisher’s method and truncated product method are much smaller when compared to the other methods.

How sensitive are these tests to unmeasured confounding? The maximum p-values for the
Table 14: Evidence factors analysis of the lead absorption study. (1) The first half of the table: Maximum p-values corresponding to the five tests with $\Gamma_k = \Gamma$, $1 \leq k \leq 5$. We dropped the subscript $a_{k-1}$ from $P_{k, a_{k-1}, \Gamma}$ used in Section 3.3–Section 6. (2) The second half of the table: Maximum p-values for testing at least $k$ of $H_0; l_s$ are false when the bias is at most $\Gamma_1 = \cdots = \Gamma_5 = \Gamma$ (using the truncated product method; truncation level $\xi = 0.20$). The maximum p-values less than 0.05 are highlighted in bold.

<table>
<thead>
<tr>
<th>$\Gamma$</th>
<th>$P_{5,\Gamma}$</th>
<th>$P_{4,\Gamma}$</th>
<th>$P_{3,\Gamma}$</th>
<th>$P_{2,\Gamma}$</th>
<th>$P_{1,\Gamma}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.420036</td>
<td><strong>0.009441</strong></td>
<td>0.095923</td>
<td><strong>0.00381</strong></td>
<td><strong>0.00007</strong></td>
</tr>
<tr>
<td>1.2</td>
<td>0.470253</td>
<td><strong>0.013512</strong></td>
<td>0.128619</td>
<td><strong>0.006773</strong></td>
<td><strong>0.00263</strong></td>
</tr>
<tr>
<td>1.4</td>
<td>0.512934</td>
<td><strong>0.017814</strong></td>
<td>0.161157</td>
<td><strong>0.010557</strong></td>
<td><strong>0.00688</strong></td>
</tr>
<tr>
<td>1.6</td>
<td>0.549884</td>
<td><strong>0.022219</strong></td>
<td>0.192914</td>
<td><strong>0.015089</strong></td>
<td><strong>0.001425</strong></td>
</tr>
<tr>
<td>1.8</td>
<td>0.582428</td>
<td><strong>0.02672</strong></td>
<td>0.223553</td>
<td><strong>0.020268</strong></td>
<td><strong>0.002525</strong></td>
</tr>
<tr>
<td>2</td>
<td>0.611224</td>
<td><strong>0.031257</strong></td>
<td>0.252909</td>
<td><strong>0.025994</strong></td>
<td><strong>0.004007</strong></td>
</tr>
<tr>
<td>2.2</td>
<td>0.636902</td>
<td><strong>0.035769</strong></td>
<td>0.280914</td>
<td><strong>0.032177</strong></td>
<td><strong>0.005867</strong></td>
</tr>
<tr>
<td>2.4</td>
<td>0.659949</td>
<td><strong>0.040228</strong></td>
<td>0.307565</td>
<td><strong>0.038738</strong></td>
<td><strong>0.008085</strong></td>
</tr>
<tr>
<td>2.6</td>
<td>0.680756</td>
<td><strong>0.044615</strong></td>
<td>0.332889</td>
<td><strong>0.045607</strong></td>
<td><strong>0.010632</strong></td>
</tr>
<tr>
<td>2.8</td>
<td>0.699635</td>
<td><strong>0.049196</strong></td>
<td>0.356935</td>
<td><strong>0.052721</strong></td>
<td><strong>0.013472</strong></td>
</tr>
<tr>
<td>3</td>
<td>0.716841</td>
<td><strong>0.053123</strong></td>
<td>0.379764</td>
<td><strong>0.060029</strong></td>
<td><strong>0.016569</strong></td>
</tr>
<tr>
<td>4.6</td>
<td>0.813707</td>
<td><strong>0.083478</strong></td>
<td>0.526577</td>
<td><strong>0.122381</strong></td>
<td><strong>0.046885</strong></td>
</tr>
<tr>
<td>4.8</td>
<td>0.822295</td>
<td><strong>0.08707</strong></td>
<td>0.541509</td>
<td><strong>0.130282</strong></td>
<td><strong>0.051015</strong></td>
</tr>
<tr>
<td>5</td>
<td>0.830333</td>
<td><strong>0.090589</strong></td>
<td>0.555832</td>
<td><strong>0.138152</strong></td>
<td><strong>0.055166</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$\Gamma$</th>
<th>$P_5^{[1]}$</th>
<th>$P_4^{[1]}$</th>
<th>$P_3^{[1]}$</th>
<th>$P_2^{[1]}$</th>
<th>$P_1^{[1]}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.193477</td>
<td><strong>0.017172</strong></td>
<td>0.000795</td>
<td><strong>0.00002</strong></td>
<td><strong>0.00002</strong></td>
</tr>
<tr>
<td>1.2</td>
<td>0.24579</td>
<td><strong>0.027005</strong></td>
<td>0.001965</td>
<td><strong>0.00012</strong></td>
<td><strong>0.00012</strong></td>
</tr>
<tr>
<td>1.4</td>
<td>0.297852</td>
<td><strong>0.037873</strong></td>
<td>0.003864</td>
<td><strong>0.00052</strong></td>
<td><strong>0.00052</strong></td>
</tr>
<tr>
<td>1.6</td>
<td>0.348663</td>
<td><strong>0.049288</strong></td>
<td>0.006544</td>
<td><strong>0.0016</strong></td>
<td><strong>0.0016</strong></td>
</tr>
<tr>
<td>1.8</td>
<td>0.149304</td>
<td><strong>0.026378</strong></td>
<td>0.001114</td>
<td><strong>0.00204</strong></td>
<td><strong>0.00204</strong></td>
</tr>
<tr>
<td>2</td>
<td>0.161532</td>
<td><strong>0.033879</strong></td>
<td>0.002024</td>
<td><strong>0.003305</strong></td>
<td><strong>0.003305</strong></td>
</tr>
<tr>
<td>2.2</td>
<td>0.17212</td>
<td><strong>0.041805</strong></td>
<td>0.004979</td>
<td><strong>0.007052</strong></td>
<td><strong>0.007052</strong></td>
</tr>
<tr>
<td>2.4</td>
<td>0.181238</td>
<td>0.050012</td>
<td><strong>0.009511</strong></td>
<td><strong>0.012336</strong></td>
<td><strong>0.012336</strong></td>
</tr>
<tr>
<td>2.6</td>
<td>0.191565</td>
<td>0.05838</td>
<td><strong>0.030932</strong></td>
<td><strong>0.030932</strong></td>
<td><strong>0.030932</strong></td>
</tr>
<tr>
<td>2.8</td>
<td>0.205224</td>
<td>0.066817</td>
<td><strong>0.049496</strong></td>
<td><strong>0.049496</strong></td>
<td><strong>0.049496</strong></td>
</tr>
<tr>
<td>3</td>
<td>0.219255</td>
<td>0.075254</td>
<td><strong>0.054454</strong></td>
<td><strong>0.054454</strong></td>
<td><strong>0.054454</strong></td>
</tr>
<tr>
<td>4</td>
<td>0.293328</td>
<td>0.116672</td>
<td><strong>0.054454</strong></td>
<td><strong>0.054454</strong></td>
<td><strong>0.054454</strong></td>
</tr>
<tr>
<td>4.8</td>
<td>0.354142</td>
<td>0.148886</td>
<td><strong>0.054454</strong></td>
<td><strong>0.054454</strong></td>
<td><strong>0.054454</strong></td>
</tr>
<tr>
<td>5</td>
<td>0.369251</td>
<td>0.156729</td>
<td><strong>0.054454</strong></td>
<td><strong>0.054454</strong></td>
<td><strong>0.054454</strong></td>
</tr>
</tbody>
</table>

five tests are presented at the top half of Table 14. At significance level 0.05, of the five tests, the first, second, and the fourth test rejects the corresponding hypotheses, assuming no bias from unmeasured confounding. These tests become sensitive at bias levels $\Gamma_1 = 4.8$, $\Gamma_2 = 2.8$, and $\Gamma_4 = 3$, respectively. The type-I error is at most 0.05 in each column. Across
the rows the type-I error is not controlled in this top half of the Table 14. If we control for the type-I error using Bonferroni correction, we would compare the maximum p-values to $0.05/5 = 0.01$. The first test becomes sensitive at $\Gamma_1 = 2.6$, the second test at $\Gamma_2 = 1.4$ and the fourth test is sensitive even at $\Gamma_4 = 1.2$.

The bottom half of Table 14 presents a sensitivity analysis for the partial conjunctions of the tests. By Corollary 3, this part of the table provides an adaptive analysis, in the sense that the total type-I error is at most 0.05. Learning from the results of the simulation study in §6.6 we chose the truncated product method with truncation level 0.20 in computing the partial conjunction p-values. When the bias is at most $\Gamma = 1.5$ we have evidence to reject at least 3 of the 5 basic nulls. When the bias is at most $\Gamma = 2$ we no longer have evidence to reject 3, but the evidence allows us to reject 2 of the 5 basic nulls.

6.8. Conclusion
Study of a causal hypothesis is enhanced when directed tests are considered for the various predictions of the hypothesis. Of course, these testable predictions of a causal hypothesis would be based on acknowledged theories at the time when the causal hypothesis is being investigated. Inherent to these predictions are requirements of simplicity and falsifiability.

On the other spectrum of etiology, a statistical analysis of a causal or etiologic hypothesis should focus on comprehensive reports that help explicate the step from an observed data to corroboration of the hypothesis. With this aim, this chapter presents a method of analysis of an elaborate theory of predictions of a causal hypothesis. We consider such elaborate theory whose falsifiable statements can be set up as alternative hypotheses in statistical hypothesis testing problems. An etiologic hypothesis can still be false because some other prediction of the hypothesis is not true. But the focus of this chapter has been to assess the extent of which the observed data supports the predictions in the elaborate theory. Our analysis suggests decomposing the tests of the elaborate theory into nearly independent factors. Partial conjunctions of these tests tell us about fractions of the elaborate theory.
As the tests might themselves be biased by unmeasured confounding, we also consider a multi-parameter sensitivity analysis. We are thus able to quantify the bias levels at which the observed data supports a certain fraction of the elaborate theory. When the tools of this analysis are appropriately chosen, the overall type-I error of this analysis is controlled without having to pay a price for having considered multiple tests, thus, without losing any power.
A.1. Comparison of combining methods

The power of different methods of combination functions has not been evaluated previously in the context of sensitivity analysis of evidence factors. Our simulation is based on the structure of the Life Span Study data. There are $S$ strata of triplets with exposures zero-dose, low-dose and high-dose; one of each. Response is Bernoulli with probability $\expit(\alpha_s)$ if exposed zero-dose or $\expit(\alpha_s + \beta_l)$ if exposed to low-dose or $\expit(\alpha_s + \beta_h)$ for high-dose; here $\expit(x) = \exp(x)/(1 + \exp(x))$ and $\beta_l \leq \beta_h$. The strata effect $\alpha_s$ is sampled independently from $N(0, 0.2^2)$ and we consider $\beta_l \leq \beta_h$.

Two evidence factors are, one from comparing low-dose units to high-dose units and the second one from the comparison of zero-dose units to low-dose and high-dose units. In this simulation scenario we are working under the favorable situation where there is no bias in treatment assignment due to unmeasured confounders in either of the factors. The justification for choosing the favorable situation for the power computation is explained in Hansen et al. (2014, §3). We provide a brief discussion of this choice here. Even though in practice we cannot know if we are in the favorable situation, by computing the power in this situation we are assessing the ability of our analyses to discriminate between two situations where we know unambiguously the desired result of the sensitivity analyses. In one situation, with moderate bias and no treatment effect, we expect that any associations between treatment and outcome for the two evidence factors can be explained by magnitudes of bias at most $\Gamma_1, \Gamma_2$, and by construction there can be a risk of at most $\alpha$ to report otherwise. In the second situation, when there is no bias and there is a treatment effect, then we hope to reject the null hypothesis. On the other hand, if we were considering a situation where there were large biases in treatment assignments and a small treatment effect, so that rejection of the null is nearly assured for all small or moderate $\Gamma_1, \Gamma_2$, then we would not have been pleased to reject $H_0$ for small or moderate $\Gamma_1, \Gamma_2$ because we know we would also have rejected $H_0$ in this situation had it been true.
Table 15: Power of sensitivity analysis for evidence factors for Fisher’s combination method and the truncated product method based on $S = 200$ strata. Estimation of power is based on 1000 iterations. $\Gamma = \Gamma_1 = \Gamma_2$.

<table>
<thead>
<tr>
<th>$\beta_l$</th>
<th>$\beta_h$</th>
<th>$\Gamma = 1$</th>
<th>$\Gamma = 2$</th>
<th>$\Gamma = 5$</th>
<th>$\Gamma = 8$</th>
<th>$\Gamma = 10$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Fisher</td>
<td>tP</td>
<td>Fisher</td>
<td>tP</td>
<td>Fisher</td>
</tr>
<tr>
<td>0 0</td>
<td>0 0</td>
<td>0.048</td>
<td>0.045</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0 1</td>
<td>1 1</td>
<td>0.146</td>
<td>0.169</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0 2</td>
<td>1 1</td>
<td>0.999</td>
<td>0.091</td>
<td>0.13</td>
<td>0.001</td>
<td>0.001</td>
</tr>
<tr>
<td>0 3</td>
<td>1 1</td>
<td>1 1</td>
<td>0.905</td>
<td>0.935</td>
<td>0.348</td>
<td>0.433</td>
</tr>
<tr>
<td>0 4</td>
<td>1 1</td>
<td>1 1</td>
<td>1 1</td>
<td>0.922</td>
<td>0.956</td>
<td>0.749</td>
</tr>
<tr>
<td>1 1</td>
<td>1 1</td>
<td>0.215</td>
<td>0.261</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1 2</td>
<td>1 1</td>
<td>0.973</td>
<td>0.966</td>
<td>0.001</td>
<td>0.001</td>
<td>0</td>
</tr>
<tr>
<td>1 3</td>
<td>1 1</td>
<td>1 1</td>
<td>0.142</td>
<td>0.142</td>
<td>0.001</td>
<td>0.002</td>
</tr>
<tr>
<td>1 4</td>
<td>1 1</td>
<td>1 1</td>
<td>1 1</td>
<td>0.807</td>
<td>0.785</td>
<td>0.128</td>
</tr>
<tr>
<td>2 2</td>
<td>1 1</td>
<td>1 1</td>
<td>1 1</td>
<td>0.188</td>
<td>0.25</td>
<td>0.001</td>
</tr>
<tr>
<td>2 3</td>
<td>1 1</td>
<td>1 1</td>
<td>0.768</td>
<td>0.811</td>
<td>0.094</td>
<td>0.123</td>
</tr>
<tr>
<td>2 4</td>
<td>1 1</td>
<td>1 1</td>
<td>1 1</td>
<td>0.975</td>
<td>0.98</td>
<td>0.34</td>
</tr>
<tr>
<td>3 3</td>
<td>1 1</td>
<td>1 1</td>
<td>1 1</td>
<td>0.999</td>
<td>1</td>
<td>0.734</td>
</tr>
<tr>
<td>3 4</td>
<td>1 1</td>
<td>1 1</td>
<td>1 1</td>
<td>1</td>
<td>0.964</td>
<td>0.975</td>
</tr>
<tr>
<td>4 4</td>
<td>1 1</td>
<td>1 1</td>
<td>1 1</td>
<td>1</td>
<td>0.998</td>
<td>0.999</td>
</tr>
</tbody>
</table>

Fisher, Fisher’s combination method; tP, truncated product method with $\bar{\alpha} = 0.20$.

Results for the sensitivity analysis are given in Table 15, for $S = 200$, and Table 16, for $S = 500$. From the simulated power reported in the tables, we see that, there is no one combining method among the two which is uniformly better than the other. But, the truncated product method has more power compared to Fisher’s combination method for most of the cases. In a few cases when both $\beta_l$, effect of low-dose, and $\beta_h - \beta_l$, excess amount of effect of high-dose over low-dose, are large, for moderate values of $\Gamma$ Fisher’s method is slightly less sensitive. This can be explained as in such scenarios both the factors would contribute toward Fisher’s combination even though they may not have very strong evidences separately. But, as $\Gamma$ increases the truncated product method dominates Fisher’s method. All the $p$-values are calculated using the mh function of the R package sensitivity2x2xk (Rosenbaum and Small, 2015).
Table 16: Power of sensitivity analysis for evidence factors for Fisher’s combination method and the truncated product method based on $S = 500$ strata. Estimation of power is based on 1000 iterations. $\Gamma = \Gamma_1 = \Gamma_2$.

<table>
<thead>
<tr>
<th>$\beta_1$</th>
<th>$\beta_2$</th>
<th>$\Gamma = 1$ Fisher</th>
<th>$\Gamma = 1$ tP</th>
<th>$\Gamma = 4$ Fisher</th>
<th>$\Gamma = 4$ tP</th>
<th>$\Gamma = 5$ Fisher</th>
<th>$\Gamma = 5$ tP</th>
<th>$\Gamma = 9$ Fisher</th>
<th>$\Gamma = 9$ tP</th>
<th>$\Gamma = 10$ Fisher</th>
<th>$\Gamma = 10$ tP</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 0</td>
<td></td>
<td>0.045</td>
<td>0.051</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>0 1</td>
<td>1 1</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>0 2</td>
<td>1 1</td>
<td>0.85</td>
<td>0.882</td>
<td>0.353</td>
<td>0.431</td>
<td>0.002</td>
<td>0.002</td>
<td>0.002</td>
<td>0.002</td>
<td>0.002</td>
<td>0.002</td>
</tr>
<tr>
<td>0 3</td>
<td>1 1</td>
<td>1 1 1</td>
<td>1 1 1</td>
<td>0.732</td>
<td>0.788</td>
<td>0.535</td>
<td>0.606</td>
<td>0.606</td>
<td>0.606</td>
<td>0.606</td>
<td>0.606</td>
</tr>
<tr>
<td>0 4</td>
<td>1 1</td>
<td>1 1 1</td>
<td>1 1 1</td>
<td>1 1 1</td>
<td>1 1 1</td>
<td>0.998</td>
<td>0.999</td>
<td>0.999</td>
<td>0.999</td>
<td>0.999</td>
<td>0.999</td>
</tr>
<tr>
<td>0 5</td>
<td>1 1 1</td>
<td>1 1 1</td>
<td>1 1 1</td>
<td>1 1 1</td>
<td>1 1 1</td>
<td>1 1 1</td>
<td>1 1 1</td>
<td>1 1 1</td>
<td>1 1 1</td>
<td>1 1 1</td>
<td>1 1 1</td>
</tr>
<tr>
<td>1 1</td>
<td></td>
<td>0 0</td>
<td>0.0</td>
<td>0 0</td>
<td>0 0</td>
<td>0 0</td>
<td>0 0</td>
<td>0 0</td>
<td>0 0</td>
<td>0 0</td>
<td>0 0</td>
</tr>
<tr>
<td>1 2</td>
<td>1 1</td>
<td>0.018</td>
<td>0.023</td>
<td>0 0</td>
<td>0 0</td>
<td>0 0</td>
<td>0 0</td>
<td>0 0</td>
<td>0 0</td>
<td>0 0</td>
<td>0 0</td>
</tr>
<tr>
<td>1 3</td>
<td>1 1</td>
<td>0.938</td>
<td>0.925</td>
<td>0.343</td>
<td>0.332</td>
<td>0 0</td>
<td>0 0</td>
<td>0 0</td>
<td>0 0</td>
<td>0 0</td>
<td>0 0</td>
</tr>
<tr>
<td>1 4</td>
<td>1 1</td>
<td>1 1 1</td>
<td>1 1 1</td>
<td>1 1 1</td>
<td>1 1 1</td>
<td>0.998</td>
<td>0.4 0.466</td>
<td>0.266</td>
<td>0.331</td>
<td>0.965</td>
<td>0.98 0.915</td>
</tr>
<tr>
<td>1 5</td>
<td>1 1 1</td>
<td>1 1 1</td>
<td>1 1 1</td>
<td>1 1 1</td>
<td>1 1 1</td>
<td>1 1 1</td>
<td>1 1 1</td>
<td>1 1 1</td>
<td>1 1 1</td>
<td>1 1 1</td>
<td>1 1 1</td>
</tr>
<tr>
<td>2 2</td>
<td>1 1</td>
<td>0.98</td>
<td>0.988</td>
<td>0.598</td>
<td>0.663</td>
<td>0 0</td>
<td>0 0</td>
<td>0 0</td>
<td>0 0</td>
<td>0 0</td>
<td>0 0</td>
</tr>
<tr>
<td>2 3</td>
<td>1 1 1</td>
<td>1 1 1</td>
<td>1 1 1</td>
<td>1 1 1</td>
<td>1 1 1</td>
<td>0.096</td>
<td>0.123</td>
<td>0.022</td>
<td>0.036</td>
<td>0.48 0.478</td>
<td>0.223 0.237</td>
</tr>
<tr>
<td>2 4</td>
<td>1 1 1</td>
<td>1 1 1</td>
<td>1 1 1</td>
<td>1 1 1</td>
<td>1 1 1</td>
<td>0.899</td>
<td>0.891</td>
<td>0.712</td>
<td>0.69</td>
<td>0.899</td>
<td>0.891 0.912</td>
</tr>
<tr>
<td>2 5</td>
<td>1 1 1</td>
<td>1 1 1</td>
<td>1 1 1</td>
<td>1 1 1</td>
<td>1 1 1</td>
<td>0.975</td>
<td>0.983</td>
<td>0.882</td>
<td>0.912</td>
<td>0.975</td>
<td>0.983 0.912</td>
</tr>
<tr>
<td>3 3</td>
<td>1 1 1</td>
<td>1 1 1</td>
<td>1 1 1</td>
<td>1 1 1</td>
<td>1 1 1</td>
<td>0.998</td>
<td>0.998</td>
<td>0.998</td>
<td>0.999</td>
<td>0.998</td>
<td>0.998 0.999</td>
</tr>
<tr>
<td>3 4</td>
<td>1 1 1</td>
<td>1 1 1</td>
<td>1 1 1</td>
<td>1 1 1</td>
<td>1 1 1</td>
<td>1 1 1</td>
<td>1 1 1</td>
<td>1 1 1</td>
<td>1 1 1</td>
<td>1 1 1</td>
<td>1 1 1</td>
</tr>
<tr>
<td>3 5</td>
<td>1 1 1</td>
<td>1 1 1</td>
<td>1 1 1</td>
<td>1 1 1</td>
<td>1 1 1</td>
<td>1 1 1</td>
<td>1 1 1</td>
<td>1 1 1</td>
<td>1 1 1</td>
<td>1 1 1</td>
<td>1 1 1</td>
</tr>
</tbody>
</table>

Fisher, Fisher’s combination method; tP, truncated product method with $\tilde{\alpha} = 0.20$. 
A.2. Proofs

A.2.1. Proofs of Corollary 1 and Theorem 4

of Corollary 1. Let \((\bar{\Gamma}_1, \bar{\Gamma}_2)\) be the minimum point of the set \(G \cap \{\Gamma_1 \mid \Gamma_1 \geq \bar{\Gamma}_1\} \times \{\Gamma_2 \mid \Gamma_2 \geq \bar{\Gamma}_2\}\). If this set is empty there is nothing to prove, otherwise existence of \((\bar{\Gamma}_1, \bar{\Gamma}_2)\) is ensured by the assumptions on \(G\) that it is continuous and intersects with at least one boundary. If \(G\) is the full grid then \((\bar{\Gamma}_1, \bar{\Gamma}_2) = (\bar{\Gamma}_1, \bar{\Gamma}_2)\). If \(H_0\) is false there is nothing to prove. Consider \(H_0\) is true. Any false rejection, by monotonicity, would imply rejection at \((\bar{\Gamma}_1, \bar{\Gamma}_2)\). Thus probability of any false rejection is at most \(\Pr(E_{\bar{\Gamma}_1, \bar{\Gamma}_2} < \alpha) \leq \alpha\), since \(E_{\bar{\Gamma}_1, \bar{\Gamma}_2} \geq \text{Unif}[0,1]\). ■

of Theorem 4. Let \(\alpha' = \exp(-\chi^2_{4,1-\alpha}/2)\), where \(\chi^2_{4,1-\alpha}\) denotes that \((1-\alpha)\)th quantile of \(\chi^2_4\) distribution. Then, the null is rejected at significance level \(\alpha\) if and only if \(E_{1,\Gamma_1|n}E_{2,\Gamma_2|n} < \alpha'\).

Proof of (6): Since \(\{(x, y) \mid xy < \alpha'\}\) is a decreasing subset of \([0, 1]^2\) by (4)

\[
\Pr(E_{1,\Gamma_1|n}E_{2,\Gamma_2|n} < \alpha') \leq \Pr(\tilde{E}_{1,\Gamma_1|n}\tilde{E}_{2,\Gamma_2|n} < \alpha') \leq \Pr(\tilde{E}_{1,\Gamma_1|n}\tilde{E}_{2,\Gamma_2|n} \leq \alpha').
\]

By exercise 4.2.7 of [Dembo & Zeitouni (2010)], which says that the rate in the product space adds up, because \(\{(x, y) \mid xy \leq \alpha'\}\) is a compact subset of \([0, 1]^2\)

\[
\limsup_{n \to \infty} \frac{1}{n} \log \Pr(E_{1,\Gamma_2|n} < \alpha') \leq - \inf_{(x,y) \mid xy \leq \alpha'} (I_{1,\Gamma_1}(x) + I_{2,\Gamma_2}(y)) \\
\leq - \inf_{(x,y) \mid x \leq \alpha \text{ or } y \leq \alpha} (I_{1,\Gamma_1}(x) + I_{2,\Gamma_2}(y)) \leq - \max_{j=1,2} \inf_{x \mid x \leq \alpha} I_{j,\Gamma_j}(x).
\]

The second inequality follows from the fact that \(\alpha' < \alpha\) hence \(\{(x, y) \mid x \leq \alpha \text{ or } y \leq \alpha\}\) contains \(\{(x, y) \mid xy \leq \alpha'\}\) and the final inequality is true since \(I_{j,\Gamma_j}\) is non-negative.
Proof of (7):

\[
\Pr(E_{1,1}; E_{2,1}, E_{2,2}|n \geq \alpha') \leq 1 - \Pr(E_{1,1}|n < \alpha'^{1/2}, E_{2,2}|n < \alpha'^{1/2}) \\
= 1 - \Pr(E_{1,1}|n < \alpha'^{1/2}) - \Pr(E_{2,2}|n < \alpha'^{1/2}) + \Pr(\{E_{1,1}|n < \alpha'^{1/2}\} \cup \{E_{2,2}|n < \alpha'^{1/2}\}) \\
\leq 1 - \Pr(\tilde{E}_{1,1}|n < \alpha'^{1/2}) - \Pr(\tilde{E}_{2,2}|n < \alpha'^{1/2}) + \Pr(\{\tilde{E}_{1,1}|n < \alpha'^{1/2}\} \cup \{\tilde{E}_{2,2}|n < \alpha'^{1/2}\}) \\
= 1 - \Pr(\tilde{E}_{1,1}|n < \alpha'^{1/2}, \tilde{E}_{2,2}|n < \alpha'^{1/2}) \\
= \Pr(\{\tilde{E}_{1,1} \geq \alpha'^{1/2}\} \cup \{\tilde{E}_{1,1} \geq \alpha'^{1/2}\}).
\]

Where the inequality follows from Definition 2 as \{((x, y) \mid x \leq \alpha'^{1/2}, y \leq \alpha'^{1/2}\} is a decreasing set. We write

\[
\limsup_{n \to \infty} \frac{1}{n} \log \Pr(E_{1,1,2}|n \geq \alpha) \leq - \inf_{(x, y)|x \geq \alpha^{1/2} \text{ or } y \geq \alpha^{1/2}} (I_{1,1}(x) + I_{2,2}(y)) \\
\leq - \inf_{(x, y)|x \geq \alpha} (I_{1,1}(x) + I_{2,2}(y)) \leq - \max_{j=1,2} \inf_{x|\alpha \geq \alpha} I_{j,1}(x).
\]

The first inequality uses the fact that rate in the product space adds up and the second inequality follows since \alpha'^{1/2} \geq \alpha for \alpha < 0.20.

\[
\]

\[\]

A.2.2. Proof that \((E_{1,1}, E_{2,2})\) in the Life Span Study data analysis are evidence factors

Let \(E_{1,1}\) be the evidence, i.e. maximum \(p\)-value, for the proximal survivors with high doses versus low doses and let \(E_{2,2}\) be the evidence, i.e. maximum \(p\)-value, for the proximal survivors versus not-in-city residents. We shall prove that \((E_{1,1}, E_{2,2})\) form evidence factors as per Definition 2. We make the following observations.

Remark 16 \(E_{2,2}\) is marginally stochastically larger than uniform random variable on \([0, 1]\). For a proof of this see Rosenbaum (2002, §4.4).

Remark 17 Treatment assignment for the two comparisons can be thought of as a two stage process. In the first stage participants get assigned to the proximal survivor group.
In the next stage proximal survivors get assigned to high doses group. Let $z_f$ and $z_s$ be the first and second stage treatment assignments respectively. Since $E_{2,\Gamma_2}$ uses the first stage treatment assignment while the evidence $E_{1,\Gamma_1}$ calculated conditional on the first stage assignment,

$$
Pr(E_{1,\Gamma_1} \leq p_2 \mid E_{2,\Gamma_2}) = E(Pr(E_{1,\Gamma_1} \leq p_2 \mid z_f) \mid E_{2,\Gamma_2}) \leq E(p_2 \mid E_{2,\Gamma_2}) = p_2.
$$

Where the inequality again follows from the same argument as in Remark 16.

**Remark 18** Going back to the notation of Definition 2, let $U_1$ and $U_2$ be two independent uniformly distributed random variables on $[0, 1]$. We use the theory of Shaked and Shanthikumar (2007, §6-B). Since $(U_1, U_2)$ is an independent pair it is a conditionally increasing in sequence.

By Remark 16, $E_{2,\Gamma_2} \succ U_1$ and by Remark 17, $E_{1,\Gamma_1} \mid E_{2,\Gamma_2} \succ U_2$. Thus using Theorem 6.B.4 of Shaked and Shanthikumar (2007, §6-B) we finally get, $(E_{1,\Gamma_1}, E_{2,\Gamma_2}) \succ (U_1, U_2)$. Thus the proof is complete.

**A.2.3. The intuition behind the proposed evidence factors in the Life Span Study**

Both factors use the proximal survivors. Yet the design provides two comparisons which are nearly independent. We give some intuition behind this by considering two related examples.

Consider testing the equality of means of three groups with equal and known variance. Then two comparisons consists of (a) comparing the group 1 and 2 to group 3, (b) comparing group 1 and 2. As in our design both comparisons use the first two groups. It is a simple algebra that the two t-test statistics for the two comparisons are uncorrelated.

This is not a feature particular to t-statistics. Let’s consider testing independence of two rows in a $2 \times 3$ contingency table. To test for this hypothesis we can consider two tests on two $2 \times 2$ contingency tables. (a) The $2 \times 2$ table of the first two columns and two rows, (b)
the $2 \times 2$ table of where the first column is the sum of the first two columns in the $2 \times 3$

\[ \text{table and the second column is the third column of the original table. Let the two } \chi^2 \text{ test}
\]

\[ \text{statistics be } c_{(a)} \text{ and } c_{(b)} \text{ each with a single degree of freedom. Then, there is the Pearson's}
\]

\[ \chi^2 \text{-statistics test statistics, say } c_P, \text{ with two degrees of freedom. } c_{(a)} \text{ and } c_{(b)} \text{ are not quite}
\]

\[ \text{uncorrelated, as in the t-statistics, but nearly independent in the sense of the discussion of}
\]

\[ \S2.3 \text{ of the main text. Under the null, } (c_{(a)}, c_{(b)}) \text{ is stochastically larger than } c_P.
\]

\[ \text{Also see } \text{Alam (1974) for a simple example of two independent rank tests, they are called}
\]

\[ \text{sequential ranks in that chapter, in testing for independence in three groups.}
\]

A.3. Algorithm for detection of the retention set

\[ \text{For this description consider } d \text{ evidence factors with biases } \{ \Gamma_{1,i} \mid i = 1, \ldots, l_1 \}, \ldots, \{ \Gamma_{d,i} \mid
\]

\[ i = 1, \ldots, l_d \} \text{ with } E_{j\Gamma_j, i} \text{ denoting the evidence for } j\text{th evidence factor and bias level } \Gamma_{j,i}
\]

\[ \text{for index } i \text{ running from } 1, \ldots, l_j \text{ and } j \text{ running from } 1, \ldots, d. \text{ When the combination}
\]

\[ \text{method is increasing in all its coordinates under the monotonicity condition (2) of the main}
\]

\[ \text{text, the retention set is convex increasing subset as described in } \S2.5.1 \text{ of the chapter.}
\]

\[ \text{The retention set is uniquely identified by its lower boundary. The following algorithm runs in } \Omega(d \log \max_j l_j)
\]

\[ \text{time. The following pseudo code } \text{retentionBrd} \text{ uses familiar R (R Development Core Team, 2012)
\]

\[ \text{concepts for simplicity of illustration and returns the}
\]

\[ \text{boundary of the retention set.}
\]

\[ \text{Algorithm 1 Function: retentionBrd. Input: } d, \{ \Gamma_{j,i} \}, \{ E_{j,i} \}.
\]

\[ \text{Initialize } B = \text{as.list}(\Gamma_{1,i} \mid i = 1, \ldots, l_1); \text{ count } = 1
\]

\[ \text{While count } < d
\]

\[ \text{For gamIdx } = 1 \text{ to gamIdx } = l_1
\]

\[ \text{Set projGam } = B[[\text{gamIdx}]]
\]

\[ \text{If length(projGam) } == \text{ count}
\]

\[ \text{temp } = \text{findBorder}(d, \{ \Gamma_{j,i} \}, \{ E_{j,i} \}, \text{projGam}, \text{ count } + 1)
\]

\[ B[[\text{gamIdx}]] = c(B[[\text{gamIdx}]], \text{temp})
\]
count = count + 1

Return B

Algorithm 2 Function: findBorder. Input: \( d, \{\Gamma_{j,i}\}, \{E_{j,i}\}, \text{projGam}, dIdx. \)

Initialize \( Evnew = 1 \)

For \( j = 1 \) to \( j = \text{length}(\text{projGam}) \)

\[ Evnew = Evnew \times E_{j,\text{projGam}[j]} \]

\[ Evnew = Evnew \times \{E_{dIdx,i} \mid i = 1,\ldots,l_{dIdx}\} \times \prod_{j=dIdx+1}^{d} E_{j,t_j} \]

Initialize \( i1 = 1; i2 = l_{dIdx} \)

If rejected for \( d \) evidences \( Evnew[i2] \)

Return \( \Gamma_{dIdx,i2} \)

If NOT rejected for \( d \) evidences \( Evnew[i1] \)

Return NULL

While TRUE

If \( |i2 - i1| <= 1 \)

index = i1; break;

itemp = \((i1 + i2)/2\)

If NOT rejected for \( d \) evidences \( Evnew[ceiling(itemp)] \)

\( i2 = ceiling(itemp) \)

If rejected for \( d \) evidences \( Evnew[floor(itemp)] \)

\( i1 = floor(itemp) \)

Return \( \Gamma_{dIdx,index} \)

A.4. Amplification of the sensitivity parameter \( \Gamma \)

This section presents an amplification of the sensitivity parameter \( \Gamma \), as defined in Section 3.2 of the main text, using two parameters which relate the unmeasured confounding to the response and the treatment assignment. For a detailed discussion on the topic see [Rosenbaum and Silber (2009)](#).
Let $i = 1, \ldots, n$ be the indices assigned arbitrarily to $n$ units. Let $Z_i$ be the indicator for unit $i$ being in the treatment group. Also, let $x_i$ denote the observed pretreatment covariates while $u_i$ is an unobserved number summarizing the unobserved confounders for unit $i$. Finally, suppose unit $i$ if exposed to treatment would have response $r_{Ti}$ and if spared exposure would have response $r_{Ci}$. Let $\mathcal{F} = \{(r_{Ti}, r_{Ci}, x_i, u_i) \mid i = 1, \ldots, n\}$. Then the sensitivity parameter $\Gamma$ is defined by, 

$$
\Gamma = \max_{1 \leq i, i' \leq n}(\Pr(Z_i = 1 \mid \mathcal{F})\Pr(Z_{i'} = 0 \mid \mathcal{F})\Pr(Z_{i'} = 1 \mid \mathcal{F})\Pr(Z_i = 0 \mid \mathcal{F}))^{-1} | x_i = x_{i'}).
$$

As the above definition conditions on $\mathcal{F}$ which includes the potential responses, it inherently assumes a near perfect relationship between the response and $u$ (Rosenbaum, 2002, §4).

Let $\mathcal{C} = \{(x_i, u_i) \mid i = 1, \ldots, n\}$. We can model the influence of $u$ on $r$ and $Z$ separately on a principal stratification that conditions on $\mathcal{C}$ rather than on $\mathcal{F}$. We assume there are $S$ strata of two units one exposed one unexposed, the $j$th unit in strata $s$ be randomly denoted by $s_j$, $j = 1, 2$, $s = 1, \ldots, S$. Then, consider a parameter $\Delta$ so that

$$
\Pr((r_{Cs_1} - r_{Cs_2}) > y \mid \mathcal{C}) = \exp(\log(\Delta)(u_{s_1} - u_{s_2})) \Pr((r_{Cs_1} - r_{Cs_2}) < -y \mid \mathcal{C}).
$$

The above model tilts the unit with larger value value of $u$ towards a larger value of $r_C$. As for the influence of $u$ on the treatment assignment, we consider a model of similar form

$$
\Pr(Z_{s_1} = 1, Z_{s_2} = 0 \mid \mathcal{C}) = \frac{\exp(\log(\Delta)u_{s_1})}{\exp(\log(\Delta)u_{s_1}) + \exp(\log(\Delta)u_{s_2})} = 1 - \Pr(Z_{s_1} = 0, Z_{s_2} = 1 \mid \mathcal{C}).
$$

The larger the value of $\Delta$ and $\Lambda$ the larger the bias. The correspondence of the sensitivity parameter $\Gamma$ to the two bias parameters $\Delta$ and $\Lambda$ is given by $\Gamma = (\Delta \Lambda + 1)/(\Delta + \Lambda)$ (see Rosenbaum and Silber, 2009, Proposition 1). More specifically, if one calculates the maximum $p$-value assuming $\Delta$ and $\Lambda$ as the two numbers determining the relationship of $u$ with the treatment and the response respectively, this $p$-value is the same as the maximum $p$-value calculated by under sensitivity parameter value $\Gamma = (\Delta \Lambda + 1)/(\Delta + \Lambda)$ and conditioning on $\mathcal{F}$. Figure 11 shows the amplification curves for various values of $\Gamma$. Observe the duality.
Amplification curves of \((\Delta, \Lambda)\) for various \(\Gamma\) values

Figure 11: Correspondence of the sensitivity parameter \(\Gamma\) with the pair of parameters \((\Delta, \Lambda)\).

relation – as \(\Delta \to \infty\), \(\Lambda\) asymptotes to \(\Gamma\), similarly as \(\Lambda \to \infty\), \(\Delta\) asymptotes to \(\Gamma\).
Proof of Lemma 3. First we note that $T_{nm}$ is a function of $Y_{n\{s\}}$ which are simply linear functions of $Z_{n\{si\}}$. Given the strata, from equation (4.5) we have that the maximum p-value of the narrow versus marginal comparison, $P_{nm,\Gamma_{nm}}$, is computed based on the conditional distributions $\{[Z_{n\{si\}} | \sum_{i\in[n]} Z_{n\{si\}} + \sum_{i\in[m]} Z_{m\{si\}}]\}$. Combining these facts, we get the first result that marginally $P_{nm,\Gamma_{nm}}$ is a function of $\{Z_{n\{si\}}\}$ and $\sum_{i\in[n]} Z_{n\{si\}} + \sum_{i\in[m]} Z_{m\{si\}}$.

Next we note that $T_{bc}$ is a function of $\sum_{i\in[n]} Z_{n\{si\}} + \sum_{i\in[m]} Z_{m\{si\}}$. Now, by looking at equation (4.6), $P_{bc,\Gamma_{bc}}$ is computed based on the family of conditional distributions $\{[\sum_{i\in[n]} Z_{n\{si\}} + \sum_{i\in[m]} Z_{m\{si\}} | \sum_{i\in[n]} Z_{n\{si\}} + \sum_{i\in[m]} Z_{m\{si\}} + \sum_{i\in[c]} Z_{c\{si\}}]\}$. Consequently, $P_{bc,\Gamma_{bc}}$ is determined by the number of exposed cases $\{\sum_{i\in[n]} Z_{n\{si\}} + \sum_{i\in[m]} Z_{m\{si\}}\}$ and the total number of exposed individuals $\{\sum_{i\in[n]} Z_{n\{si\}} + \sum_{i\in[m]} Z_{m\{si\}} + \sum_{i\in[c]} Z_{c\{si\}}\}$. But it is enough to know whether each control is exposed or not, i.e., $Z_{c\{si\}}$, to know the number of exposed cases when we have the information on total number of exposed units. The result is hence proved. \[\blacksquare\]

Proof of Lemma 4. For part (i) and (ii) note that p-values or their upper bounds are valid p-values thus are stochastically larger than $Unif[0,1]$. Part (iii) and (iv) follows from (i) and (ii) simply by marginalizing since marginalization preserves stochastic ordering. \[\blacksquare\]

Proof of Lemma 5. Note that, since conditional on $\sum_{i\in[n]} Z_{n\{si\}} + \sum_{i\in[m]} Z_{m\{si\}}$ the random variables $Z_{n\{si\}}$ and $Z_{c\{si\}}$ are independently distributed, by Lemma 3 the conditional distribution in the statement of the lemma is same as $[P_{nm,\Gamma_{nm}} | \sum_{i\in[n]} Z_{n\{si\}} + \sum_{i\in[m]} Z_{m\{si\}}]$. Now the result follows from part (i) of Lemma 4. \[\blacksquare\]
Proof of Lemma 6. We can write for any $0 \leq p, q \leq 1$, the conditional probability as,

$$\Pr(P_2 \leq q \mid P_1 \leq p) \overset{by \ C_1}{=} \Pr(P_2 \leq q \mid \{V_1 : P_1 \leq p\})$$

$$= \mathbb{E}\left[\Pr(P_2 \leq q \mid V_1) \mid \{V_1 : P_1 \leq p\}\right]$$

$$\overset{by \ C_2}{\leq} \mathbb{E}\left[q \mid \{V_1 : P_1 \leq p\}\right] = q.$$

The second equality above follows from the tower property of conditional expectation. The lemma then follows. □
APPENDIX C : APPENDIX TO CHAPTER 6

Proof of Theorem 13. Let \( \mathcal{K} = \{k_1, \ldots, k_{|\mathcal{K}|}\} \subseteq \{1, \ldots, K\} \) and \( U_1, \ldots, U_{|\mathcal{K}|} \) be \( |\mathcal{K}| \) i.i.d. random variables uniform on \([0, 1]\). Since \( P_{k_1, a_{k_1-1}, \Gamma_{k_1}} \) is the maximum of \( P_{k_1, a_{k_1-1}} \) over the unmeasured confounders \( u_{ij_{k_1}} \)'s. For \( \alpha_1 \in [0, 1] \) we have

\[
\Pr( P_{k_1, a_{k_1-1}, \Gamma_{k_1}} \leq \alpha_1 \mid \mathcal{F}, H_{0,k_1} ) \\
\leq \Pr( P_{k_1, a_{k_1-1}} \leq \alpha_1 \mid \mathcal{F}, H_{0,k_1} ) \\
\leq E[ \Pr( P_{k_1, a_{k_1-1}} \leq \alpha_1 \mid Z_{k_1-1}, \sum_{ij \in I_{k_1-1}(a_{k_1-1})} Z_{ij}^{(k_1)} ) , \mathcal{F}, H_{0,k_1} ] \leq E[\alpha_1] = \Pr( U_1 \leq \alpha_1 ).
\]

The expectation in the previous calculation is over the \( Z_{k_1-1}, \sum_{ij \in I_{k_1-1}(a_{k_1-1})} Z_{ij}^{(k_1)} \) conditional on \( \mathcal{F}, H_{0,k_1} \). We borrow the notation of Shaked and Shanthikumar (2007). Then \( U_1 \leq \text{st } P_{k_1, a_{k_1-1}, \Gamma_{k_1}} \).

Now let \( 2 \leq l \leq |\mathcal{K}| \). Note that for any \( k_l \) the maximum p-value \( P_{k_l, a_{k_l-1}, \Gamma_{k_l}} \) is a function of \( Z_{l} \) and \( \mathcal{F} \). Hence, for \( \alpha_l \in [0, 1], \)

\[
\Pr( P_{k_l, a_{k_l-1}, \Gamma_{k_l}} \leq \alpha_l \mid P_{k_{l-1}, a_{k_{l-1}}, \Gamma_{k_{l-1}}}, \ldots, P_{k_1, a_{k_1-1}, \Gamma_{k_1}, \Gamma_{k_l-1}, \ldots, \Gamma_{k_{l-1}}}, \mathcal{F}, H_{0,k_l} ) \\
\leq \Pr( P_{k_l, a_{k_l-1}, \Gamma_{k_l}} \leq \alpha_l \mid P_{k_{l-1}, a_{k_{l-1}}, \Gamma_{k_{l-1}}}, \ldots, P_{k_1, a_{k_1-1}, \Gamma_{k_1}, \Gamma_{k_l-1}, \ldots, \Gamma_{k_{l-1}}}, \mathcal{F}, H_{0,k_l} ) \\
\leq E[ \Pr( P_{k_l, a_{k_l-1}, \Gamma_{k_l}} \leq \alpha_l \mid Z_{l-1}, \sum_{ij \in I_{k_l-1}(a_{k_l-1})} Z_{ij}^{(k_l)}, P_{k_{l-1}, a_{k_{l-1}}, \Gamma_{k_{l-1}}}, \ldots, P_{k_1, a_{k_1-1}, \Gamma_{k_1}} ) , \mathcal{F}, H_{0,k_l} ] \\
\leq E[ \Pr( P_{k_l, a_{k_l-1}, \Gamma_{k_l}} \leq \alpha_l \mid Z_{l-1}, \sum_{ij \in I_{k_l-1}(a_{k_l-1})} Z_{ij}^{(k_l)}, \mathcal{F}, H_{0,k_l} ) ] \\
\leq E[\alpha_l] = \Pr( U_l \leq \alpha_l ).
\]
Thus under $\cap_{t \in \mathcal{K}} H_{0,t}$ and conditional on $\mathcal{F}$,

$$
U_l \leq \Pr \left( \frac{\mathcal{F}_{k_1,a_{k_1-1},r_{k_1}} \leq \alpha_l}{\mathcal{F}_{k_1,a_{k_1-1},r_{k_1}}, \ldots, \mathcal{F}_{k_{l-1},a_{k_{l-1}-1},r_{k_{l-1}}}} \right)
$$

for all $2 \leq l \leq |\mathcal{K}|$.

Also, $(U_1, \ldots, U_{|\mathcal{K}|})$ is a conditionally increasing in sequence (CIS) (see, eq 6.B.11 of Shaked and Shanthikumar (2007)). Thus, by Theorem 6.B.4 of Shaked and Shanthikumar (2007) under $\cap_{t \in \mathcal{K}} H_{0,t}$ and conditional on $\mathcal{F}$, $(U_1, \ldots, U_{|\mathcal{K}|}) \leq \Pr((\mathcal{F}_{k_1,a_{k_1-1},r_{k_1}}, \ldots, \mathcal{F}_{k_{|\mathcal{K}|},a_{k_{|\mathcal{K}|}-1},r_{k_{|\mathcal{K}|}}) \in U)$. Now to complete the proof set $U = \{(x_1, \ldots, x_{|\mathcal{K}|}) : f_K(x_1, \ldots, x_{|\mathcal{K}|}) > x\}$ and note that $U$ is an upper set since $f_K$ is coordinatewise nondecreasing.

**Proof of Proposition 6.**

1. Consider first the ‘minimum p-value’ method. Condition (a) is obviously true. Next, note that $\Pr(\min\{U_k, \ldots, U_K\} \leq p) = 1 - (1 - p)^{K-k+1}$. Thus condition (b) is satisfied. Since, $\Pr(g_k(U_k, \ldots, U_K) \leq \alpha) = \Pr(\min\{U_k, \ldots, U_K\} \leq 1 - (1 - \alpha)^{1/(K-k+1)}) = 1 - (1 - (1 - \alpha)^{1/(K-k+1)})^{(K-k+1)} = \alpha$. Finally, to check condition (c) fix $x_k \leq x_{k+1} \leq \cdots \leq x_K$. To check $g_k(x_k, \ldots, x_K) \leq g_{k+1}(x_{k+1}, \ldots, x_K)$, it is enough to show that $(1 - x_k)^{K-k} - (1 - x_{k+1})^{K-k+1} \geq 0$. This is true since, $(1 - x_k)^{K-k} - (1 - x_{k+1})^{K-k+1} \geq (1 - x_{k+1})^{K-k} - (1 - x_k)^{K-k+1} = (1 - x_{k+1})^{K-k} x_{k+1} \geq 0$.

2. Proofs of condition (a) and (b) are straightforward for Stouffer’s method. To check condition (c) consider $x_k \leq x_{k+1} \leq \cdots \leq x_K$. Then, after some rearranging $g_k(x_k, \ldots, x_K) \leq g_{k+1}(x_{k+1}, \ldots, x_K)$ is equivalent to the inequality, $\sqrt{(K - k + 1)/(K - k - 1)}(\Phi^{-1}(1 - x_{k+1}) + \cdots + \Phi^{-1}(1 - x_K)) \leq \Phi^{-1}(1 - x_k)$. Since $x_k \leq \min\{x_{k+1}, \ldots, x_K\}$, it is enough to check that this condition holds with $x_k = 1 - \Phi(\Phi^{-1}(1 - x_{k+1}) + \cdots + \Phi^{-1}(1 - x_K))/(K - k)$.

Then the check reduces to checking $(\sqrt{(K - k + 1)/(K - k - 1)}(\Phi^{-1}(1 - x_{k+1}) + \cdots + \Phi^{-1}(1 - x_K))/(K - k))$. Finally, to complete the proof set $U = \{(x_1, \ldots, x_{|\mathcal{K}|}) : f_K(x_1, \ldots, x_{|\mathcal{K}|}) > x\}$ and note that $U$ is an upper set since $f_K$ is coordinatewise nondecreasing.
\[
\Phi^{-1}(1 - x_K) \leq (\Phi^{-1}(1 - x_{k+1}) + \cdots + \Phi^{-1}(1 - x_K))/(K - k), \text{ or } (\sqrt{(K - k + 1)/(K - k)} - 1) \leq 1/(K - k), \text{ or } \sqrt{1 + 1/(K - k)} \leq 1 + 1/(K - k); \text{ which is true.}
\]

3. Finally, consider the ‘modified additive p-value’ method. Condition (a) is obvious since \(g_k\) is an increasing function of \(A_k = x_k + \cdots + x_K\). For condition (b) note from Edgington (1972), \(\Pr(U_k + \cdots + U_K \leq x) \leq x^{K-k+1}/(K - k + 1)!\). Let \(F(x) := \Pr(U_k + \cdots + U_K \leq x)\). Then \(F(x) \leq \min(1, x^{K-k+1}/(K - k + 1)!) \leq \min(1, x^{K-k+1}/(K - k + 1)!)^{1/(x \leq c_k)}\). Thus, \(\Pr(\min(1, (U_k + \cdots + U_K)^{K-k+1}/(K - k + 1)!)^{1/(x \leq c_k)} \leq \alpha) \leq \Pr(F(U_k + \cdots + U_K) \leq \alpha) \leq \alpha\).

For condition (c) fix \(x_k \leq x_{k+1} \leq \cdots \leq x_K\). If \(A_{k+1} = x_{k+1} + \cdots + x_K > c_{k+1}, g_{k+1}(x_{k+1}, \ldots, x_K) = 1\), thus the condition is satisfied. Suppose now \(x_{k+1} + \cdots + x_K \leq c_{k+1}\). Clearly, \(x_k \leq (x_{k+1} + \cdots + x_K)/(K - k) = A_{k+1}/(K - k)\); thus \(g_k(x_k, \ldots, x_K) \leq g_k(A_{k+1}/(K - k), x_{k+1}, \ldots, x_K)\). Hence, suffices to show \(g_k(A_{k+1}/(K - k), x_{k+1}, \ldots, x_K) \leq g_{k+1}(x_{k+1}, \ldots, x_K)\). Note that, \(A_{k+1}/(K - k) + x_{k+1} + \cdots + x_K = A_{k+1}(K - k + 1)/(K - k)\).

Since, \(A_{k+1} \leq c_{k+1}, \text{ we get, } A_{k+1}(K - k + 1)/(K - k) \leq c_k\). Hence, by simple reduction \(g_k(A_{k+1}/(K - k), x_{k+1}, \ldots, x_K) \leq g_{k+1}(x_{k+1}, \ldots, x_K)\) is equivalent to \(A_{k+1}^{K-k+1}(K - k + 1)^{K-k}/(K - k)^{K-k+1} \leq A_{k+1}^{K-k}\); which simplifies to \(A_{k+1} \leq (K - k)(1 - 1/(K - k + 1))^{K-k} = c_{k+1}\). Thus proving condition (c).

**Sketch of proof of Corollary 2.** The proof is in line of the proof of Theorem 14 given in the main text. The main observation is that the thresholding level of the sensitivity parameter, \(\bar{\Gamma}\) exists even when \(\mathcal{J}\) is not a grid but a one dimensional hyper-plane \(\mathcal{J} = \{\bar{\Gamma} = \Gamma_l(1, \ldots, 1) : 1 = \Gamma_1 < \cdots < \Gamma_L\}\). Thus, probability of rejecting any of the true null among \(\{H^k_{0,\Gamma} : \Gamma \in \mathcal{J}\}\) is at most \(\Pr(P^k_{a,\bar{\Gamma}} \leq \alpha) \leq \alpha\).

**Sketch of proof of Corollary 3.** If there is no null among \(\{H^k_0 : 1 \leq k \leq K\}\) is true, there is nothing to prove. Otherwise, suppose \(H^1_{0,K}\) is the first one in the list which
is true. Recall that, under conditions (a)–(c) of Proposition 5, which is assumed in this corollary, for any \( \Gamma \) we have \( P_{a, \Gamma}^{1|K} \leq \cdots \leq P_{a, \Gamma}^{K|K} \). Thus, rejection of any true null in \( \{H_{0, \Gamma}^{|K|}; \Gamma \in J, 1 \leq k \leq K \} \) will mean that a true null in \( \{H_{0, \Gamma}^{|l|}; \Gamma \in J \} \) is rejected. Define \( \tilde{\Gamma} \) as in the proof of Theorem 14 or Corollary 2. Since \( P_{a, \Gamma}^{l|K} \) is nondecreasing in \( \Gamma \), rejecting any true null among \( \{H_{0, \Gamma}^{|K|}; \Gamma \in J, 1 \leq k \leq K \} \) means rejecting \( H_{0, \Gamma}^{|l|} \), which has probability at most \( \alpha \).

**Proof of Proposition 7.** Recall that \( P_{a, \Gamma}^{k|K}(g) = g_k(\overline{P}_{(k)a, \Gamma}, \ldots, \overline{P}_{(K)a, \Gamma}) \). Consider the first case, \( \Gamma_l > \tilde{\Gamma}_l \) for at most \( k \) many \( l \). It follows from the definition of design sensitivity that the largest \( K - k + 1 \) p-values converge to 1. Thus, \( P_{a, \Gamma}^{k|K}(g) \rightarrow g_k(1, \ldots, 1) = 1 \). In the second case, \( \Gamma_l < \tilde{\Gamma}_l \) for \( k \) or more \( l \)'s. By the definition of design sensitivity \( \overline{P}_{(l)a, \Gamma} \rightarrow 0 \) for \( l = 1, \ldots, k \) and the rest goes to 1. Thus \( P_{a, \Gamma}^{k|K}(g) \rightarrow g_k(0, \ldots) = 0 \).

**Proof of Theorem 15.** By the assumption, \( c(I)^{-1} \log \overline{P}_{(k, a_{k-1}, \Gamma_k} \rightarrow -r_k(\Gamma_k) \) almost surely for \( k = 1, \ldots, K \). Let \( r_{(1)} \Gamma \leq \cdots \leq r_{(K)} \Gamma \) be the ordered values of \( r_{1}(\Gamma_1), \ldots, r_{K}(\Gamma_K) \). As \( I \) increases to infinity \( c(I)^{-1} \log \overline{P}_{(l)a, \Gamma} \rightarrow -r_{(K-l+1)} \Gamma \) for \( 1 \leq l \leq K \) almost surely.

Fix \( k \). From the above we note that \( c(I)^{-1} \sum_{l=k}^{K} \log \overline{P}_{(l)a, \Gamma} \rightarrow -\sum_{l=1}^{K-k+1} r_{(l)} \Gamma \) almost surely. Choose \( a < -\sum_{l=1}^{K-k+1} r_{(l)} \Gamma < b \). We allow \( a = -\infty \) and \( -\infty < -\infty \). Consequently, for any \( \epsilon > 0 \) there exists \( I_\epsilon \) such that for \( I \geq I_\epsilon \), as \( c(I) \rightarrow \infty \) when \( I \) increases to infinity, with probability at least \( 1 - \epsilon \) we get \( a < c(I)^{-1} \sum_{j=k}^{K} \log \overline{P}_{(j)a, \Gamma} < b \). For \( I \geq I_\epsilon \) with probability at least \( 1 - \epsilon \)

\[
\Pr(\chi^2_{2(K-k+1)} > -2c(I)a) \leq \Pr(\chi^2_{2(K-k+1)} > -2 \sum_{j=k}^{K} \log \overline{P}_{(j)a, \Gamma}) \leq \Pr(\chi^2_{2(K-k+1)} > -2c(I)b).
\]
Noting that, \( \lim_{n \to \infty} n^{-1} \log \Pr(\chi^2_d > nx) = -x/2 \) for any \( x \geq 0 \) and \( d > 0 \) we get

\[
a \leq \liminf_{I \to \infty} c(I)^{-1} \log \Pr(\chi^2_d(2K-k+1) > -2\sum_{l=k}^{K} \log \mathcal{P}(l)_{a,\Gamma})
\]

\[
\leq \limsup_{I \to \infty} c(I)^{-1} \log \Pr(\chi^2_d(2K-k+1) > -2\sum_{l=k}^{K} \log \mathcal{P}(l)_{a,\Gamma}) \leq b.
\]

This is true for arbitrary \( \epsilon > 0 \) and arbitrary numbers \( a \) and \( b \) such that \( a < -\sum_{l=1}^{K-k+1} \Gamma(\l) \Gamma < b \). Thus we conclude that

\[
\lim_{I \to \infty} c(I)^{-1} \log P_{a,\Gamma}^{k|K}(\epsilon f) = \lim_{I \to \infty} c(I)^{-1} \log \Pr(\chi^2_d(2K-k+1) > -2\sum_{l=k}^{K} \log \mathcal{P}(l)_{a,\Gamma})
\]

\[
= -\sum_{l=1}^{K-k+1} \Gamma(\l).
\]

This limit might be negative infinity.

Now consider \( \log P_{a,\Gamma}^{k|K}(g) = \log g_k(\mathcal{P}(k)_{a,\Gamma}, \ldots, \mathcal{P}(K)_{a,\Gamma}) \) for any \( g \). From the assumption of the theorem we have \( \Pr(g_k(U_k, \ldots, U_K) \leq \alpha) \leq \alpha \), for any \( \alpha \in [0,1] \). Thus for any \( 0 \leq x_k \leq \cdots \leq x_K \)

\[
g(x_k, \ldots, x_K) \geq \Pr(g_k(U_1, \ldots, U_{K-k+1}) \leq g_k(x_k, \ldots, x_K)).
\]

By the nondecreasing property of the function \( g_k \)

\[
\Pr(g_k(U_1, \ldots, U_{K-k+1}) \leq g_k(x_k, \ldots, x_K)) \geq \Pr(U_1 \leq x_k, \ldots, U_K \leq x_K)
\]

\[
= \prod_{l=1}^{K-k+1} \Pr(U_l \leq x_{j+k-1}) = \prod_{l=k}^{K} x_l.
\]

Thus, \( g(x_k, \ldots, x_K) \geq \prod_{l=k}^{K} x_l \). This implies

\[
\log P_{a,\Gamma}^{k|K}(g) \geq \sum_{l=k}^{K} \log \mathcal{P}(l)_{a,\Gamma}.
\]
We get by dividing by \( c(I) \) and taking the limit, for \( 1 \leq k' \leq k \)

\[
\liminf_{I \to \infty} c(I)^{-1} \log P_{a,\Gamma}^{k'|K}(\varrho) \geq \lim_{I \to \infty} c(I)^{-1} \sum_{l=k}^{K} \log P_{(l)\varrho}^{(l)}
\]

\[
= - \sum_{l=1}^{K-k+1} r(l)\varrho \geq - \sum_{l=1}^{K-k'+1} r(l)\varrho = \lim_{I \to \infty} c(I)^{-1} \log P_{a,\Gamma}^{k'|K}(\varrho).
\]

**Proof of Proposition 8.** Following the proof of Theorem 15 we have, for any \( k = 1, \ldots, K \),
\[
\lim_{I \to \infty} c(I)^{-1} \log P_{a,\Gamma}^{k|K} = - \sum_{l=1}^{K-k+1} r(l)\varrho. \text{ Consider } k \text{ such that } \Gamma_k > \tilde{\Gamma}_k. \text{ Since, } \tilde{\Gamma}_k \text{ is the design sensitivity of the } k \text{th factor, by definition of the design sensitivity, } P_{k,a_k-\Gamma_k} \to 1.
\]
Further, since \( c(I) \to \infty \) as \( I \to \infty \), it implies \( r_k(\Gamma_k) = 0 \). The number of \( l \) with \( \Gamma_l < \tilde{\Gamma}_l \) is called \( \tilde{k} \). Hence, in the ordered values \( r(1)\varrho \leq \cdots \leq r(K)\varrho \) the first \( K - \tilde{k} \) are zero. Thus the proof of the first part follows.

To prove of the final statement, consider the truncated product method. Let \( x \) be the truncation level. For a number \( a \) let \( a^x \) be the truncated version defined as \( a \) if \( a < x \), otherwise it is 1. The combining method is \( g_k(x_k, \ldots, x_K) = \Pr(\prod_{l=k}^{K} U_l^x < \prod_{l=k}^{K} x_l^x) \), where \( U_1, \ldots, U_K \) are i.i.d. uniform(0,1) random variables. For \( 0 \leq x_1, \ldots, x_K \leq 1 \), let \( y = -c(I)^{-1} \log \prod_{l=k}^{K} x_l^x \). We write with \( I \to \infty \) in mind (and \( K \) fixed)

\[
g_k(x_k, \ldots, x_K)
= \Pr(\prod_{l=k}^{K} U_l^x < \exp(-c(I)y/2))
= \Pr(-2 \sum_{l=k}^{K} \log U_l^x > c(I)y)
= \sum_{K \subseteq \{k, \ldots, K\}} \Pr(-2 \sum_{l=k}^{K} \log U_l^x > c(I)y \mid U_j \geq x, \forall j \in K^c) \Pr(U_j \geq x, \forall j \in K^c)
= \sum_{K \subseteq \{k, \ldots, K\}} \Pr(-2 \sum_{l=k}^{K} \log U_l > c(I)y \mid U_j \geq x, \forall j \in K^c) \Pr(U_j \geq x, \forall j \in K^c)
\]
\[
\sum_{K \subseteq \{k, \ldots, K\}} \Pr(-2 \sum_{i \in K} \log U_i > c(I)y) \Pr(U_j \geq x, \forall j \in K')
\]
\[
= \sum_{K \subseteq \{k, \ldots, K\}, K \neq \emptyset} \Pr(\chi_{2|K|}^2 > c(I)y) \times (1 - x)^{|K'|}
\]
\[
= \sum_{K \subseteq \{k, \ldots, K\}, K \neq \emptyset} \exp\left\{ -c(I)y/2 + o(c(I)) \right\} \times (1 - x)^{|K'|}
\]
\[
= \exp\left\{ -c(I)y/2 + o(c(I)) \right\} \sum_{K \subseteq \{k, \ldots, K\}, K \neq \emptyset} (1 - x)^{|K'|}
\]
\[
= \exp\left\{ -c(I)y/2 + o(c(I)) \right\} \times \{1 - (1 - x)^{K-k+1}\}.
\]

We used the fact that \(\lim_{n \to \infty} n^{-1} \log \Pr(\chi_d^2 > nx) = -x/2\) for any \(x \geq 0\) and \(d > 0\). Using the truncated product method (call it \(\text{tp}\))

\[
P^{k|K}_{a, \Gamma}(\text{tp}) = \exp\left\{ -\log \prod_{l=k}^{K} P^{x}_{(l), a, \Gamma} + o(c(I)) \right\} \times \{1 - (1 - x)^{K-k+1}\}.
\]

Thus, \(c(I)^{-1} \log P^{k|K}_{a, \Gamma}(\text{tp}) = \{-\sum_{l=k}^{K} c(I)^{-1} \log P^{x}_{(l), a, \Gamma} + o(1)\} + o(1)\). Finally, for large \(I\), \(P^{x}_{(l), a, \Gamma} = P^{x}_{(l), a, \Gamma}\) for all \(l\) since \(P^{x}_{(l), a, \Gamma}\) converges to 0 or 1, in this setting. We get, from our proof of Theorem 15, \(c(I)^{-1} \log P^{k|K}_{a, \Gamma}(\text{tp}) - c(I)^{-1} \log P^{x|K}_{a, \Gamma}(\text{ef}) = o(1)\). This completes the proof.
BIBLIOGRAPHY


Karmakar, B. (2017). approxmatch: Approximately optimal fine balance matching with


Zubizarreta, J. R., Neuman, M., Silber, J. H. and Rosenbaum, P. R. (2012). Contrasting evidence within and between institutions that provide treatment in an observational study