Exhibit 293

EPIDEMIOLOGY AND THE LAW

David A. Savitz, PhD
and Stephen G. Schwarz, Esq.

JURIS

Questions about This Publication

For assistance with shipments, billing or other customer service matters, please contact the Customer Service Department at:

customerservice@jurispub.com

For questions of an editorial nature please contact the Editorial Department at

editorial@jurispub.com

To obtain a copy of this book, or to find other law-related titles please visit:

www.jurispub.com

Printed in the United States of America.

ISBN 978-1-57823-563-6

Juris Publishing, Inc.
www.jurispub.com

ABOUT THE BOOK

The field of epidemiology, the scientific discipline in which patterns of exposure and disease in human populations are described ("descriptive epidemiology") and used to make inferences about the causes of disease ("analytic epidemiology") has become more important each year to lawyers practicing in toxic tort, pharmaceutical product liability and a number of other areas of law where issues of causation associated with exposures to a toxic substance or a drug or medical device are litigated. Understanding epidemiological studies requires familiarity with the terminology, statistical methods, and reasoning used in the field, which is foreign to most lawyers.

In writing this book, the authors, an experienced academic epidemiologist who has also acted as an expert witness and served on numerous panels of experts in assessing causes of disease, and a trial lawyer with years of experience in cases involving complex causation issues, have tried to provide lawyers with an accessible text to learn basic concepts and reasoning used in epidemiology. Additionally, the text provides a detailed review of the case law demonstrating how various state and federal courts throughout the United States have dealt with expert causation testimony in this field. Our intention is to provide a valuable reference tool for lawyers and judges who are confronted with epidemiological issues in their cases. We also believe this book will be of value to ancillary professionals working with law firms, such as legal nurse consultants or environmental health specialists, who assist in screening potential cases and educating lawyers on various scientific and medical topics. Finally, we believe this book will be valuable to professionals who have served or want to serve as expert witnesses in cases involving complex causation issues that touch upon epidemiology.

Several chapters explain how epidemiologists generate and use information from studies to reach conclusions regarding causality, describing the specific aspects of study design and conduct that determine the validity of the results. There are also chapters on specific topics that frequently arise in legal cases, including statistical significance testing, meta-analyses that summarize a series of studies, and interpreting the significance of negative studies. We then address how courts have dealt with the opinions of epidemiologists under *Daubert* and *Frye* challenges and under what circumstances one or more studies have been found to be sufficient or insufficient to provide an adequate foundation to establish causation in particular cases.

Above all, the authors have sought to make this book technically accurate but completely accessible to lawyers and judges, providing multiple hypothetical examples to explain the otherwise difficult and subtle scientific concepts. We intend this book to be a valuable reference to be kept on the shelf and available when background information is needed to address issues confronting lawyers and judges in a particular case, as well as numerous case references that may be helpful in briefing or deciding a particular issue.

ABOUT THE AUTHORS

David Savitz is Professor of Epidemiology in the Brown University School of Public Health with joint appointments as a Professor of Obstetrics and Gynecology and Pediatrics in the Alpert Medical School. From 2013–2017, Dr. Savitz served as the vice president for research at Brown University. He came to Brown in 2010 from Mount Sinai School of Medicine, where he had served as the Charles W. Bluhdorn Professor of Community and Preventive Medicine and the director of the Disease Prevention and Public Health Institute since 2006. Before that appointment, he taught and conducted research at the University of North Carolina School of Public Health and at the Department of Preventive Medicine and Biometrics at the University of Colorado School of Medicine. His epidemiologic research has addressed a wide range of topics including exposures to physical and chemical hazards in the workplace and community, health impact of exposures associated with military deployment, environmental effects of energy development, risks from environmental exposures during pregnancy, and drinking water safety. He has authored more than 400 papers in professional journals and the editor or author of four books on environmental epidemiology. He has served as president of the Society for Epidemiologic Research, the Society for Pediatric and Perinatal Epidemiologic Research, and the North American Regional Councilor for the International Epidemiological Association. Dr. Savitz is an elected member of the National Academy of Medicine, inducted in 2007, and has previously served on 14 consensus committees, 8 of which he chaired or vice-chaired. Dr. Savitz received his undergraduate training in psychology at Brandeis University, holds a master's degree in preventive medicine from The Ohio State University, and received his Ph.D. in epidemiology from the University of Pittsburgh Graduate School of Public Health.

Stephen G. Schwarz is Managing Partner at Faraci Lange, LLP in Rochester, New York and has been a trial lawyer for over 35 years. He was selected as a Fellow in the American College of Trial Lawyers in 2005, and served on the organization's Board of Regents from 2014–2018. He was selected as a member of the American Board of

Trial Advocates in 1995. He has litigated dozens of cases involving complex causation issues including medical, medical device and environmental contamination cases. Most recently, he has served as lead counsel in litigation involving PFOA exposures from contaminated drinking water in upstate New York. He has also litigated and tried multiple cases involving toxic exposures to TCE and other solvents as well as diseases caused by asbestos exposure. He has published numerous articles in professional journals including the ACTL Journal, Voir Dire Magazine, the journal published by ABOTA, Trial Magazine, a publication of the American Association for Justice and the New York State Bar Journal. He has successfully argued numerous appeals setting important precedents in environmental toxic tort law in New York and Federal appellate courts. His undergraduate degree from the University at Albany was in Biology and he thereafter earned a J.D. degree from Albany Law School.

SUMMARY TABLE OF CONTENTS

About the Book iii
About the Authors v
Summary Table of Contents vii
Table of Contents ix

Chapter One

INTRODUCTION 1

Chapter Two

STRUCTURE OF EPIDEMIOLOGIC RESEARCH 9

Chapter Three

INTERPRETING THE RESULTS OF EPIDEMIOLOGIC STUDIES.. 27

Chapter Four

REACHING JUDGMENTS BASED ON EPIDEMIOLOGIC EVIDENCE 49

Chapter Five

EVALUATING SUFFICIENCY OF EVIDENCE TO INFER A CAUSAL EFFECT 67

Chapter Six

THE USE AND MISUSE OF STATISTICAL SIGNIFICANCE TESTING 89

Chapter Seven

INTEGRATING EVIDENCE ACROSS STUDIES 107

Chapter Eight

INTERPRETING NEGATIVE STUDIES 119

Chapter Nine

DISTINCTIVE EXPERTISE OF EPIDEMIOLOGISTS AND HOW TO IDENTIFY IT 131

Chapter Ten

GENERAL AND SPECIFIC CAUSATION 145

Chapter Eleven

DAUBERT AND *FRYE* CHALLENGES AND OTHER PRE-TRIAL MOTIONS INVOLVING EPIDEMIOLOGY 159

Chapter Twelve

DEPOSITIONS, DIRECT AND CROSS-EXAMINATION OF EPIDEMIOLOGIC EXPERTS 169

GLOSSARY OF EPIDEMIOLOGIC TERMS 179

TABLE OF CONTENTS

About the Book iii
About the Authors v
Summary Table of Contents vii
Table of Contents ix

Chapter One

INTRODUCTION

Chapter Two

STRUCTURE OF EPIDEMIOLOGIC RESEARCH

I. Introduction 9
II. Strengths and Limitations of Epidemiology Relative to Other Approaches 11
III. Study Designs Used in Epidemiology 14
IV. Presentation of Results from Epidemiologic Studies 17
V. How Epidemiologic Studies Are Conducted 20

Chapter Three

INTERPRETING THE RESULTS OF EPIDEMIOLOGIC STUDIES

I. How Epidemiology Addresses Causality 27
II. Causal Inference Requires Examining Study Biases 28
III. Sources of Bias that Result in Invalid Results 30
A. Confounding 31
B. Exposure Measurement Error 38
C. Disease Measurement Error 41
D. Selection Bias from Losses to the Study Population 43
E. Random Error 45

Chapter Four

REACHING JUDGMENTS BASED ON EPIDEMIOLOGIC EVIDENCE

I. Introduction 49
II. Concept of Cause in Epidemiology 50
III. Is an Association Present? 52
IV. Does the Absence of an Association Indicate the Exposure Is Not a Cause of the Disease? 55
V. Does the Presence of an Association Indicate the Exposure Is a Cause of the Disease? 57
VI. Bradford-Hill Criteria for Inferring that an Association is Causal 58
VII. Alternative Strategy for Evaluating Causality of Associations between Exposure and Disease 61
VIII. Relative Strengths and Weaknesses of Study Designs 64

Chapter Five

EVALUATING SUFFICIENCY OF EVIDENCE TO INFER A CAUSAL EFFECT

I. Introduction 67
II. Clarity in the Question of a Causal Effect 70
III. Role of Ancillary Evidence 72
IV. Commonly Used Arguments in Support of a Judgment of Causality 75
A. Statistical Evidence of an Association 75
B. Evidence of a Dose-Response Gradient 76
C. Quality of the Studies Finding an Association 77
D. Absence of Clear Basis for Attributing the Association to Bias 78
E. Corroboration of Evidence for an Association Across Multiple Studies 79
F. Data on Time Trends or Geographic Patterns of Exposure or Disease 79
G. Biological Rationale for a Causal Effect 80
V. Commonly used Arguments in Opposition to a Causal Judgment 81
A. Statistical Uncertainty and Cherry-Picking 81

B. Absence of a Dose-Response Gradient 82
C. Cumulative Basis for Uncertainty 83
D. Identification and Documentation of Specific Biases that Create an Association 83
E. Evidence that Higher Quality Studies are Less Likely to Identify an Association 84
F. Time Trends or Geographic Patterns of Exposure or Disease Inconsistent with Causal Effect 84
G. Absence of Biological Rationale for Causal Effect 85
VI. Conclusion 85

Chapter Six

THE USE AND MISUSE OF STATISTICAL SIGNIFICANCE TESTING

I. Introduction 89
II. Formal Basis for Statistical Significance Testing 91
III. Arguments in Favor of Reliance on Statistical Significance Testing 95
IV. Arguments Against Reliance on Statistical Significance Testing 96
V. Confidence Intervals 99
VI. Statistical Significance and Causal Effects 101
VII. Recommended Approach to Assessing the Impact of Random Error 103

Chapter Seven

INTEGRATING EVIDENCE ACROSS STUDIES

I. Introduction 107
II. Rationale for Conducting Meta-Analyses 108
III. Disadvantages of Relying on Meta-Analysis to Synthesize Evidence 109
IV. Informative Strategies for Integrating Evidence Across Studies: Categorize Methods 112
V. Informative Strategies for Integrating Evidence Across Studies: Link Methods to Results 115
VI. Application of Evidence Synthesis to Legal Issues 116

Chapter Eight

INTERPRETING NEGATIVE STUDIES

I. Introduction 119
II. Statistical Significance Testing and Negative Studies 120
III. Common Reasons for Spurious Negative Results 122
A. Poor Quality Exposure Assessment 123
B. Poor Quality of Disease Measurement 124
C. Negative Confounding 125
D. Random Error 126
IV. Conclusions Regarding Negative Studies 127

Chapter Nine

DISTINCTIVE EXPERTISE OF EPIDEMIOLOGISTS AND HOW TO IDENTIFY IT

I. The Field of Epidemiology 131
II. Distinctive Features of Epidemiologic Evidence for Causal Inference 132
III. Skills Needed from an Epidemiologist 133
IV. *Daubert* Factors in Selecting an Epidemiology Expert 135
V. Epidemiology Expert versus Hired Gun 137
VI. Expectations from an Epidemiology Report 139
VII. Components of a Report 141

Chapter Ten

GENERAL AND SPECIFIC CAUSATION

I. Introduction 145
II. Evidence to Establish Causation 145
III. Required Magnitude of Association 149
IV. Interpretation of Statistical Tests and Confidence Intervals 152
V. Sufficiency of Evidence of Causation 152
VI. Conclusions 157

Chapter Eleven

DAUBERT AND *FRYE* CHALLENGES AND OTHER PRE-TRIAL MOTIONS INVOLVING EPIDEMIOLOGY

I. Introduction 159
II. *Daubert* and *Frye* 159
III. States Where *Frye* Remains the Rule 165
IV. Epidemiology and Class Certification 166

Chapter Twelve

DEPOSITIONS, DIRECT AND CROSS-EXAMINATION OF EPIDEMIOLOGIC EXPERTS

I. Introduction 169
II. Deposition Strategies 169
III. Experience with the Subject Matter 170
IV. Epidemiological Evidence 170
V. Fit 173
VI. Direct Examination 173
VII. Cross-Examination 176

GLOSSARY OF EPIDEMIOLOGIC TERMS 179

Chapter 1

INTRODUCTION

The purpose of this book is to make the science of epidemiology more accessible to lawyers and judges. This, we hope, will lead to a better understanding of epidemiological studies and the testimony of epidemiologists and facilitate more informative, scientifically grounded presentations or refutations of such evidence in court. Ultimately, more informed use of epidemiology in the legal setting will result in more just decisions. Our goals are to:

- Provide an understanding of the basic concepts of epidemiology to those who draw upon epidemiologic evidence in litigation;
- Propose strategies for making epidemiologic methods, research findings, and interpretation of such evidence accurate and accessible to judges and juries;
- Examine prototypic legal issues that draw upon epidemiologic evidence to provide guidance on the information to be elicited from experts;
- Elucidate frequent points of contention in competing interpretations of epidemiologic evidence.

As an epidemiologist who has been involved in a number of legal proceedings on a wide range of issues and an attorney who has tried numerous cases in which epidemiological evidence was crucial, we believe our combined experience from these differing perspectives will provide a unique and helpful approach to understanding epidemiology and applying this science in various legal contexts. We intend this book to be neither a legal nor scientific treatise, but rather, a practical tool for attorneys and judges to better understand and communicate in the language of epidemiology and to use these concepts in legal writing and trial presentations. Although one co-author has typically represented plaintiffs in attempting to prove a causal association in court, he has also frequently been required to refute some alternative causation theory allegedly supported by epidemiological data. Similarly, the other co-author has testified both as an expert supporting a causal association of some agents and refuting that there is sufficient support in the epidemiological literature to support a causal relationship for other agents. Thus, we have endeavored to make this volume helpful to all practitioners regardless of the camp to which they belong, as well as to judges who are required to determine issues of admissibility of epidemiological evidence.

The incorporation of complex scientific topics into legal proceedings is often fraught with confusion, difficulty, and inconsistency. Science, by its very nature, is always incomplete and evolving, with theories being proposed, tested, and either supported or refuted as additional studies are published. When supported, other scientists will attempt to replicate and build upon the findings, which again will further strengthen or weaken the overall support for the hypothesis, in this case concerning the causal effect of some exposure on a particular health condition. This process continues relentlessly with the certainty of the conclusions constantly shifting until some broad consensus is reached. Such thought evolution through the scientific process is particularly descriptive of the science of epidemiology, where the shifts in evidence are often widely publicized and stimulate controversy.

Conversely, the development of the common law relies on legal precedent with slow incremental expansion happening reluctantly as judges are inclined to follow the doctrine of stare decisis. Thus, in science, change is constant and expected, while in law, change is discouraged in favor of a system of stable rules that are clearly understood and designed to foster predictable outcomes. The simpler the rule, the easier it is to apply and the more predictable the outcome. When the parties can predict the likely outcome of a controversy before litigating it, they have greater incentive to resolve it without going through the expense of litigation. Thus, bright line rules are preferred in the common law. However, in epidemiology, bright lines are rare to nonexistent. Hence the inherent conflict between the two disciplines.

As an example, in one particular lawsuit a trial judge or appellate court may determine that agent X is capable of causing a certain adverse outcome. In another lawsuit, the judge or court may determine that there is insufficient scientific evidence to allow the issue of causation to reach a jury for exposure to agent Y. Twenty years after these decisions are rendered additional studies may provide strong evidence to the contrary on the causal relationship between exposures to both agents X & Y and the same adverse

outcomes discussed in those decisions. Yet, even though each case should be decided factually on the record before the court, lawyers on both sides of those issues in lawsuits filed decades later will cite each prior decision as legal precedent for or against causation, even though the scientific evidence has evolved markedly in the intervening period. Judges confronted with this prior precedent often become conflicted between their duty to honor the doctrine of stare decisis, which is central to their legal training, or reject prior precedent based upon the record before them which includes updated complex scientific evidence that is far outside of their expertise. This hypothetical is hardly speculative.[1] It is an inevitable consequence of the different fundamental approaches to training in law and epidemiology.

What further complicates this intersection between law and epidemiology is that these two disciplines utilize common terms such as *causation* and *causality* with widely differing meanings. As will be discussed in Chapters 3–7, in interpreting epidemiological data there is a spectrum of certainty from hypothesis, to possible association, to established association, to probable causal relationship, to a judgement of causality, which occurs over multiple studies that typically span decades and frequently never reach closure. The level of certainty required for confidently asserting causality in epidemiology is arguably analogous to the criminal burden of proof of beyond a reasonable doubt. Conversely, in an individual civil lawsuit, the issue is whether the alleged agent more probably than not caused the damages suffered, a difficult standard for epidemiology experts to address. Thus, *causation* can mean very different things within the fields of epidemiology and the law, and translating from one to the other can be difficult.

Epidemiology is the study of the patterns and determinants of disease in human populations, with the goal of understanding the causes of disease to determine needed actions to improve the health of the public. Trained epidemiologists conduct and review studies of populations first to determine whether there is evidence indicative of an association between some potentially causative agent and a human

[1] In later chapters we will discuss cases involving the drug Bendectin and whether it was capable of causing birth defects will be discussed. In many of those cases, judges cited and relied upon prior decisions addressing the admissibility of this causation evidence, without fully discussing whether the records in the two cases were precisely identical.

illness or condition. This typically requires comparing the frequency of disease in a group that has relatively elevated exposure to the frequency of disease in a group that is unexposed or has a lower level of exposure. There are different study designs to make this comparison, which will be further explained in Chapter 2. Each design and in fact each individual study has its strengths and weaknesses, all of which must be taken into account for the epidemiologist to reach conclusions.

When epidemiologists determine that those who are more highly exposed have an elevated risk of disease relative to those who are not, they need to make an informed judgment regarding whether it is likely that the exposure has in fact *caused* an elevated risk of disease. Even having observed an association, i.e., those with higher exposure having a greater risk of disease, the possibility that the association is spurious and not an accurate reflection of a causal effect needs to be carefully considered. While causality cannot be proven with absolute certainty, the field of epidemiology has developed clear principles and methodologic tools to make a reasoned, scientifically grounded judgment. By considering alternative explanations for the association, including biases and random error, and conducting analyses to address those alternative explanations, the case for a causal interpretation can be strengthened or weakened, depending on what is found.[2] The question of causality is central to epidemiology since the study of statistical associations alone without evaluating the causal significance offers no guidance for methods of preventing disease to improve public health. To argue that reducing exposure would improve health requires an assumption of a causal effect.

In epidemiology, there is a continuum of evidence that can support causal inferences. For instance, when discussing smoking and lung cancer, evidence of a causal effect is compelling, yet for many years this association was challenged with the simplistic mantra "correlation is not causation." The judgment to be made is whether the evidence of an association is or is not likely to reflect a causal impact. While scientific certainty of causality is difficult to establish with any exposure and may take decades of study to reach this level, epidemiologists are able to make informed use of available data to address questions of causality without awaiting the sort of overwhelming support for smoking as a cause of lung cancer. By considering the body of scientific evidence and interpreting it with an appreciation of the underlying methodologic strengths and limitations, reliable judgments can be made, including when a causal link is more likely than not to be present.

[2] One of the authors co-authored a previous volume devoted to practical strategies for making such inferences in a methodical, transparent, informative manner (SAVITZ DA, WELLENIUS GA, INTERPRETING EPIDEMIOLOGIC EVIDENCE: CONNECTING RESEARCH TO APPLICATIONS (New York: Oxford University Press, 2016)).

As will be discussed in Chapter 8, in epidemiology a negative study, i.e., a study that does not show an association between an exposure and a specific illness, also needs to be scrutinized for its validity to evaluate whether it provides meaningful evidence that no causal effect is present. Just as for a positive indication of an association, studies that generate an absence of association are subject to biases and random error which can generate false-negative findings, i.e., failing to find an association even when a causal effect is truly present. There is no reason to automatically accept "lack of correlation" as a clear indicator of "no causal effect" any more than to accept "presence of correlation" as a clear indicator of "causal effect present." The interpretation of either result calls for a thorough assessment. An overall assessment considers the full range of studies that provide pertinent information regardless of their results and integrates the full range of relevant studies. Negative studies may reflect insufficient statistical power to detect associations due to small populations or limited range of exposure, a particular challenge in studying rare types of cancer (discussed at length in Chapter 8). The agent being studied and its biological fate after it enters the human body can differ markedly. Some agents, such as perfluorooctanoic acid (**PFOA**) in drinking-water remain in the body for long periods of time and can be measured through blood testing. Conversely, trichloroethylene (**TCE**) a known human carcinogen, is metabolized quickly inside the body and no accurate test method for its presence once inhaled or ingested has been devised. Studies that do not measure or estimate exposure accurately are also more likely to fail to detect a true association that may be present, with the error in exposure estimation tending to shift measures of association towards the null value (showing little or no association).

Randomized clinical trials, which are often conducted to assess the effectiveness of medications, provide a methodologically sound basis for assessing causal effects, both favorable and adverse consequences. In such trials, study participants are randomly selected to receive the active treatment or no treatment (*e.g.*, placebo) and are therefore balanced by age, gender, medical history, etc. Because the treatment is assigned randomly, we can be confident that if the intervention has no effect, the groups will have a similar risk of the health outcome, referred to as having "equal baseline risk." With that features, whatever differences are found must be attributable to the drug or random error. Double-blind studies, where neither the researchers nor the subjects know who is getting the drug and who is getting the placebo, ensure freedom from biased assessments by the patients or researchers. However, when researching potentially toxic agents, epidemiologists cannot ethically conduct experiments with controls where one group of people is intentionally exposed to a suspected toxic agent while a control group is not, and then monitor these groups to compare how many from each group develops a particular disease. Epidemiologists must instead study groups that differ in exposure for reasons outside the control of the investigator, such as personal choice or geographic location. Studies are designed to assess the incidence of disease in exposed compared to an unexposed or less exposed population to determine whether those who were more highly exposed to the toxic substance have a greater risk of disease than those not exposed. Epidemiologists may study occupational exposures, where people in a particular occupation are exposed through their work to a suspected toxicant, or community exposures, which are often more difficult to study because of the challenge in measuring exposure and the possibility that some other attributes that happen to be associated with exposure are also causes of the disease, referred to as confounders. For example, if the rate of smoking happens to be higher in those who live near a hazardous waste site, even if the chemicals from the waste site have no adverse effects, those living near the site will have a higher risk of lung cancer and other tobacco-related diseases. In this case, smoking is a confounder of the relationship between proximity to the waste site and health outcomes and needs to be accounted for to isolate any true effect of living near the hazardous waste site.

Causation in the legal context is frequently divided into two components: General Causation—whether agent X is capable of causing adverse outcome Z; and Specific Causation—assuming agent X is capable of causing outcome Z, did agent X *actually* cause outcome Z for the plaintiff in the lawsuit. General causation is the

bailiwick of the epidemiologist. It is through studies evaluating exposure to agent X in groups of occupationally or environmentally exposed people that an epidemiologist looks for associations between exposure and various outcomes including diseases and biological markers. If an association is demonstrated and particularly if the association is more common with higher exposure, referred to a dose-response relationship, the suggestion of a causal relationship is strengthened. However, the epidemiologist must look to other factors that might affect this relationship other than just agent X to be confident that the exposure is responsible for the increased risk of disease. These other factors, referred to as *confounders*, may or may not be ruled out as impactful, depending on the design of the study. Multiple studies showing the same dose-response relationship will provide a strengthening of the association between agent X and outcome Y moving the needle in the direction of general causation. Yet, there is no bright-line test to determine where on the spectrum of certainty an association becomes generally accepted as a probable cause and effect.

The gatekeeping function of the judge under *Daubert* or *Frye* in the above example requires the trial judge to determine whether there is enough support in the scientific literature to allow a jury to conclude that a plaintiff's exposure to agent X caused outcome Y. This requires attorneys and their epidemiology experts to explain this complex literature to the court and why it either supports general causation or is insufficient for general causation to be found by a jury. If that hurdle is cleared, then the attorneys must convince a jury of laypersons whether these same epidemiological studies support a causative relationship between agent X and outcome Y. These are daunting tasks given the complexity of the science and need to explain it clearly to the court and a jury.

Anyone who has suffered from insomnia and tuned into late night television has seen a proliferation of advertising claims that various products or drugs that are alleged to have caused a variety of ailments for which compensation may be available. These can range from definitive cause and effect, such as between exposure to asbestos and the development of mesothelioma, the universally fatal cancer of the pleural lining of the lung, to much more speculative associations that have recently been suggested, such as use of hair care products and development of uterine cancer. Even in the asbestos example, there is nuance and potential controversy with regard to whether certain types of asbestos fibers (*e.g.*, chrysotile fibers) released from certain types of products (*e.g.*, brake linings) are capable of causing mesothelioma. The number and types of cases in which epidemiology has become not just relevant, but essential, seems to be growing exponentially each year.

Chapter 2

STRUCTURE OF EPIDEMIOLOGIC RESEARCH

In this chapter we will explain the basic features of epidemiologic research, the way it attempts to assess cause and effect relationships, how epidemiology compares to other scientific methods addressing health effects, study designs commonly used in epidemiology, how study results are presented, and how studies are conducted.

I. INTRODUCTION

Epidemiology is the scientific discipline in which patterns of exposure and disease in human populations are described ("descriptive epidemiology") and used to make inferences about the causes of disease ("analytic epidemiology"). Descriptive epidemiology often includes the frequency of disease occurrence based on person, place, and time, *e.g.*, cancer incidence by age or ethnicity, across counties within a state, and increases or decreases over time. Analytic epidemiology seeks to identify specific causes, *e.g.*, whether diet or pollution exposure modifies risk of disease. Just describing the frequency and pattern of disease occurrence can be useful to plan needed health services, establish priorities for investigation, and inform policy makers and the public about how disease rates differ across geographic areas, over time, and in relation to social and demographic characteristics such as age, sex, ethnicity, and social class. Such information provides the context for epidemiologic studies of potential causes of disease, clinical trials, and laboratory research into causes and treatment of disease. Descriptive epidemiology is often useful to get the "lay of the land" before embarking on a deeper look at the issues of more direct concern. It can also provide clues to possible causes of disease, for example, if an area in which some known or possible environmental exposure is present has a higher incidence of disease than other areas where that exposure is absent, or if disease is increasing over time following introduction of a new drug. Conceptually, thinking of location or ethnicity or calendar time as a "cause" of disease is too abstract or indirect to be helpful, but they can be important predictors of disease or help to see whether the patterns are supportive of a hypothesized cause. For example, if geographic areas near a chemical plant have a higher risk of certain cancers, that could help to make the case that those chemicals are causing the cancer, or if a newly introduced drug is accompanied by subsequent increases in a disease in the subsequent time period, that would support a possible causal effect.

But the more direct way to use epidemiology to address cause and effect relationships calls for a methodical, strategic evaluation of the patterns of disease in relation to putative causal agents. As described in more detail in later sections of the book, epidemiologists conduct studies to evaluate whether some agent (toxin, medication, exposure) has altered the risk of disease, either preventing it or increasing its occurrence. These indicators of exposure-disease association come from observations of patterns in human populations. It is important to distinguish these observational studies from experiments, such as those that are conducted on laboratory animals or in clinical trials of human populations. The key difference is whether the potential cause of disease is assigned randomly by the investigator, *e.g.*, drug trials, or occurs on its own outside the control of the researcher, *e.g.*, use of a consumer product. Both lines of research are intended to provide relevant information to determine whether a possible cause is in fact influencing disease risk, but the strengths and limitations of these approaches differ considerably. For some putative causes of disease, such as potential environmental carcinogens, randomized trials are not feasible so only observational studies can be used, whereas for other exposures, most notably pharmaceutical agents, both randomized controlled studies and observational research may contribute to our understanding of the effects of exposure on health outcomes. Where it is feasible to conduct both experiments and observational studies, the information can be complementary for drawing conclusions. Drug trials have notable strengths in avoiding confounding due to random exposure assignment, but they are often relatively small and of short duration, limiting the ability to identify rare, delayed adverse effects. In contrast, studies based on large healthcare databases may allow for the evaluation among those who did and did not use the drug over periods of many years among tens of thousands of individuals. The two lines of research each have strengths and limitations but differ from one another in what those strengths and limitations are.

II STRENGTHS AND LIMITATIONS OF EPIDEMIOLOGY RELATIVE TO OTHER APPROACHES

The power of epidemiologic studies is their firm grounding in the real world and thus their ability to zero in on the question of ultimate interest. Studying free-living human populations who have experienced varying levels of the potential disease determinants of interest directly addresses the driving question—has this exposure caused an increased risk of disease? There is no need to extrapolate from laboratory animals to humans or from highly selected human populations in clinical trials to the general population of interest—the exposure of concern and population of interest is studied directly. Similarly, we do not have to extrapolate from the high doses used in laboratory experiments or even the prescribed exposures in clinical trials to the ones we are really interested in—we are studying the levels of exposures that we want to know about.

Not surprisingly, there are also significant limitations in studying free-living populations that need to be recognized and mitigated to the extent possible in the design, execution, and analysis of such studies, and considered in the interpretation of the study's findings. Multiple factors that could be influencing the incidence of a disease occur together, referred to as "confounding," making it unclear which is really culpable. These factors must be distinguished, based on how the studies are designed and the methods utilized to analyze the data. Carrying matches or a lighter is undoubtedly strongly associated with the risk of developing lung cancer, but it is obvious that the real culprit is cigarette smoking. Match-carrying is associated with, but not causing, lung cancer.

It can be difficult to accurately determine the presence of the exposures and diseases we are interested in for a number of reasons. The operational definition of "exposed" and "diseased" in a given study has to be based on the information that is available. We often rely on people's memory or records of where they lived or worked to assess exposure, or other imperfect approaches since we have not assigned the exposure that they receive (as we do in a clinical trial) or carefully monitored their experiences on an ongoing basis over their lifetime. Many health conditions are difficult to ascertain accurately in an epidemiologic study, particularly if diagnoses require invasive medical procedures, constant surveillance or engaging in discretionary health care. The data that are ultimately analyzed in an epidemiologic study are referred to as exposure and disease, but inevitably, they are approximations of the exact exposure that really occurred and the presence of disease subject to being able to determine its presence or absence in each of the study participants. There is a need to ask how closely these operational definitions of exposure and disease correspond to the actual level of exposure and incidence of the disease of interest.

Finally, we need to recognize that the people we are able to include in our studies are those who are willing to enroll or otherwise may be selected based on such factors as where they get their medical care. Because they are independent human populations, not laboratory animals, there is always some degree of unintentional selection into and out of the studies that needs to be taken into account. Depending on the pattern of enrollment and retention in the study, the measure of association may be distorted and not accurately reflect the causal effect, referred to as "selection bias." For example, in a study of dietary supplements to prevent colds, if those who use the product and were free of colds are especially motivated to participate in the study, more so than those who used the supplement and nonetheless were sick, the results would make the supplement look more beneficial than it actually was.

When the same question can be addressed using different methods, namely randomized trials and observational studies, the information each type of research provides can be complementary in drawing conclusions about a causal effect. Randomized clinical trials have the notable strength of studying humans in much more controlled circumstances. In a randomized trial, the exposure of interest, often a drug thought to have potential benefit, is randomly assigned and then those who receive the drug and a placebo are monitored over time for their health outcomes. Investigators impose the exposure randomly by assigning who gets the drug and who gets the placebo rather than just observing the exposure as it occurs naturally. Random assignment helps to isolate the exposure of interest from other disease influences since we can be sure that independent of any effect of the drug, their disease risk would have been the same, referred to as "equal baseline risk." Clinical trials can monitor the health of study populations methodically on an ongoing basis rather than relying on health care records. But there are considerable restrictions on when this approach can be used. Obviously, ethical considerations preclude intentionally exposing people to potentially harmful agents. The demands of such trials often result in a highly selected population such that results may not apply directly to the much more heterogeneous population of ultimate interest. The range of exposure that is subject to manipulation may be far narrower than what people are exposed to naturally. For example, when we try to modify diet or exercise in a randomized trial, we are typically able to make subtle shifts, but left to their own devices, people manifest huge differences that can be considered in observational studies, well beyond the range of what can be imposed.

Laboratory research using experimental animals or biological materials is an even more extreme contrast to observational epidemiology. The level of control and ability to manipulate exposure has fewer ethical bounds (none for non-living material), the exposure levels can be extremely high, invasive biological examinations can be performed, and genetically homogeneous strains of laboratory animals can be used. But the information value of these tightly controlled, potentially definitive assessments is severely limited by the need to extrapolate the information back to the species (humans) and exposure conditions (those in the real world) of interest. If we were ultimately interested in whether an agent can cause cancer in genetically homogeneous rodents, this sort of research could put the issue to rest once and for all, but that assessment is just a means to the real goal of assessing what happens to humans who are exposed to levels of the agent that commonly occur.

Interpreting the multiple research approaches applicable to inferring causality requires an appreciation of the strengths and weaknesses of each. Often the practitioners of one approach or another tend to be advocates for their line of work, exaggerating the strengths and understate the limitations inherent in the methods. An informed and balanced assessment of the contributions of epidemiology is needed to use it optimally to advance medicine and public health, and to have it be appropriately understood and accurately appreciated in the legal setting.

III. STUDY DESIGNS USED IN EPIDEMIOLOGY

There are a number of research strategies used in the conduct of epidemiologic research, with three basic study designs: cohort studies, case-control studies, and ecologic studies. Cohort studies are the most analogous to experiments in which the researcher identifies people with varying degrees of exposure, referred to as the study cohort. Their health is monitored over time to determine who develops the disease of concern and who does not. What we are interested in is whether those who have greater exposure to the agent develop disease more frequently than those who have less or no exposure. These studies can be done in real time, starting by identifying the people with varying levels of exposure and following them forward into the future. Alternatively, we can use historical data to identify the cohort at some point in the past and assess their disease risk in the subsequent period which ended at some point in the past. The former is called a prospective cohort study and the latter a retrospective cohort study. The advantage of starting from scratch is that you can collect whatever exposure data you are interested in with as much detail as you want, and as you follow the cohort through time, you can actively monitor their health experience. In contrast, historical cohort studies are dependent on having accurate information on both exposures and disease occurrence in the past, relying on memory or records. The obvious advantage of a retrospective cohort study is that it is efficient—the researchers do not have to age with their cohort as is the case with a prospective cohort study, generating results more rapidly.

The main drawback to cohort studies is that they can be expensive, take a long time to yield results, and therefore can be inefficient. When the health condition of concern is rare, such as certain cancers or birth defects, we may need to study tens of thousands of people in a cohort study to generate adequately precise results, since very few people will go on to develop the health outcome we are interested in. This is less of a concern for more common diseases or for subclinical health problems such as changes in biological markers that can be readily measured (*e.g.*, cholesterol, hormones) or other health measures that fall along a continuum (*e.g.*, blood pressure, IQ) such that everyone has a value.

When studying rare health outcomes that are present or absent, it may be preferable to conduct case-control studies. In this design we assemble a group of people who have developed the disease of concern (cases) and a comparison group from the same or similar population that produced the cases (controls) to assess their history of exposure to the agent of concern. By directly seeking and engaging those with disease, we can obtain sufficient numbers through such sources as hospitals, clinics, or disease registries. Controls are selected from the same population as the cases came from (*e.g.*, the same geographic area), sometimes matched on age, sex, and other characteristics to help with the comparisons to the cases. The goal is to choose people who would have been identified as cases had they developed the disease of interest. Once we have identified the cases of disease and controls (those to whom they will be compared), we determine the exposure history of each group to find out if cases with disease are more likely to have been exposed than controls. If those with disease are more likely to have a history of exposure, a positive association is present whereas if the groups have a similar likelihood of having been exposed, there is no association present.

The main advantage of case-control studies is their efficiency in assembling a large group of cases without having to monitor a huge population over an extended period of time. For rare diseases this design is often the most widely used. The main drawbacks include challenge in finding suitable controls for comparing exposure histories and having exposure information obtained after disease has developed leading to potential recall bias, which is discussed in more detail in Chapter 3.

Frequently, a third design is used to characterize groups when we do not have information on individuals within those groups. These are referred to as ecologic studies. Disease maps are frequently used in which disease rates are indicated across census tracts, counties, or other geographic units. For instance, imagine that the incidence of breast cancer in County A is found to be twice as high as the incidence of breast cancer in neighboring County B. If we also have information on exposure to some agent in those counties, for example, air pollution levels, we may ask whether there are higher levels of air pollution in the area with higher rates of breast cancer. While intuitively appealing, with rare exceptions, this is a weak strategy for inferring cause and effect relationships and is usually a method only for determining whether more rigorous studies are warranted. This is true for several reasons: First, there is usually no information on people moving into or out of the areas. For instance, if looking for specific cancers from a cancer registry, you will find the residence address of the person when they were diagnosed. You would not find information on who moved out of the area just prior to diagnosis or those who was diagnosed after living in the area for only a brief period. Second, exposure information is often indirect and may well not apply equally to all the people within the geographic unit, i.e., some of those in the high exposure area may be heavily exposed while others are not exposed at all. Third, the time from onset of exposure to occurrence of disease cannot be measured. Many cancers and other chronic diseases have latency periods between exposure and diagnosis and in this type of study there is no way to evaluate whether a plausible latency is present. Finally, variation in health care access and utilization may drive differences in identified disease (a confounder unrelated to the exposure). The fundamental social, cultural, and economic differences among areas may be impossible to fully control. While it may seem intuitively appealing to simply compare rates of disease across areas to determine where there are specific causes, there is a great deal of random variation, with some areas having higher rates than others due to chance alone. A long history of such research by public health agencies and others has yielded very little knowledge regarding the causes of disease and is unlikely to do so in the future.

Somewhat related to ecologic studies are the reports of patterns of disease in a neighborhood or community, often generated by the residents or activists concerned with an environmental health threat. These are colloquially referred to as "cancer clusters" but have also been identified for other health problems such as miscarriages or birth defects. There may be a tally of persons suffering from health problems plotted in relation to a source of pollution of concern in order to see whether there appears to be an excess of disease among those at greatest risk of exposure. While this can suggest possibilities to investigate using more rigorous research methods, it rarely provides useful information on a potential causal effect on its own. There is often selective awareness regarding the occurrence of disease and often the health problems noted are quite heterogeneous in their potential etiology.

Beyond the range of epidemiologic studies, there may be reports on the possible connection between exposure and disease that fall into the category of anecdotes rather than research. Case reports are often published when an individual with a health problem

reports some exposure that is thought to be a possible cause of that problem. If both the exposure and disease are quite rare, for example, a rarely used medication taken during pregnancy associated with a rare birth defect in the child, it may be a clue that is worthy of pursuit through research but in itself, not informative about cause-and-effect relationships. For more common exposures or health outcomes, the value as a clue will be negligible since the co-occurrence of the two is not unexpected in a large population. For example, reporting that an individual who applied a household pesticide went on to develop migraine headaches would be of little value given a common exposure and common health problem.

A case series may also be reported in which a number of individuals with a health problem are identified as having shared some exposure of concern. Again, depending on the rarity of the exposure and the health outcome, this can be an informative clue that warrants research of a more methodical nature, such as a case-control study. In a sense, these case series constitute half of a case-control study, having assembled a group of cases and ascertained their exposure history but lacking controls to compare them to in order to determine whether the prevalence of past exposure among cases is greater than would be expected.

IV. PRESENTATION OF RESULTS FROM EPIDEMIOLOGIC STUDIES

The product of epidemiologic studies is generally a measure of the association between the potential cause of disease and the frequency of developing the disease, *e.g.*, a ratio of the disease frequency in those exposed compared to those not exposed as a relative risk or odds ratio, or a subtraction of the rate of disease among those who were not exposed from the rate among those who were exposed, referred to as a "risk difference." The goal is to provide a quantitative estimate of how much, if any, the risk of disease appears to be influenced by the exposure of concern. Note that studies do not simply determine whether there is "any association" versus "no association," but rather, provide a quantitative estimate of the association that can range from indicating complete protection against disease (relative risk of zero) to an infinite increase in risk of developing disease (relative risk of infinity) or no association at all (relative risk of 1.0). We may ultimately interpret the results as a dichotomy, i.e., the study does or does not indicate an association is present, but the study itself provides a quantitative estimate of the association. For example, in a study of benzene exposure and leukemia, we may find a relative risk of 2.7 with a 95% confidence interval of 1.3 to 5.0. We could simply note the study was "positive," which is accurate, but more information is provided by the quantitative result, which indicates a sizable association based on a sufficiently large study to result in a narrow confidence interval.

It is also important to note that studies do not directly generate measures of causal effects of exposure on disease. They only provide measures of association that are subject to interpretation regarding the extent to which the measures of association accurately approximate the true causal effect. Causality must be inferred based on research, drawing on an understanding of the details of the research that has been done and a command of epidemiologic methods. As discussed below, there are methodologic principles that guide this interpretation, but to be accurate, one cannot say that a study shows (let alone proves) that exposure causes increased risk of disease. The study shows an association; the interpreter of that study infers a causal effect and that interpretation may or may not be accurate.

The most common way that results are presented is in some form of relative risk which compares the frequency of disease occurrence among the subset of the population with higher exposure to the frequency of disease occurrence among the subset of the population with lower exposure. If 10% of those exposed to a particular agent such as a drug or chemical develop a health problem but only 5% of those not exposed to the agent develop a health problem, we would calculate the relative risk as 2.0 (10% divided by 5%), suggesting the agent may double the risk of disease. If equal numbers developed disease in both groups, say 5%, the relative risk would be 1.0 (5% divided by 5%), sometimes referred to as the null value indicating no association is present. If the risk of developing disease is lower among those exposed to the agent, say 2% compared to 5% among those not exposed to the agent, the relative risk would be 0.4 (2% divided by 5%), suggesting the agent may prevent the disease.

There are a number of specific terms used to describe these ratios of disease risk among the exposed versus the unexposed, depending on the exact ways in which the frequency of the disease was measured and the study design used to generate the measure of association. These include relative risk, risk ratio, odds ratio, rate ratio, hazard ratio, prevalence ratio, mortality ratio and others. While there is a technical basis for distinguishing these measures from one another, for interpretation they are all providing essentially the same information: an estimate of the ratio of disease risk among those who were exposed (or more highly exposed) to the disease risk among those who were not exposed (or who had lower exposure). The only distinction that is worth noting is whether the ratio is based on new cases of disease that occurred over some period of time, referred to as incidence, or is based on those with the disease present at a point in time, disease prevalence. Sometimes studies present prevalence ratios and that number depends not only on who developed disease (incidence) but on how long the disease lasts and whether the disease is fatal such that some who did develop the disease are not counted as prevalent cases.

Some features of the relative risk measures should be kept in mind. This measure does not tell us anything about how common the disease is overall—the relative risk would be 2.0 whether it is based on 10% divided by 5%, 80% divided by 40%, or 0.02% divided by 0.01%. For that reason, we sometimes present risk differences rather than ratios, subtracting the risk of disease among those unexposed from the risk of disease among those exposed. In the previous series of examples, the risk differences would be 5% (10% minus 5%), 40% (80% minus 40%), and 0.01% (0.02% minus 0.01%), making clear that doubling a very common disease has more public health impact than doubling a rare disease. Even if this calculation is not made, it is important to keep in mind the overall frequency of disease occurrence since the absolute number of people potentially affected is much greater for common diseases than for rarer diseases. The ratio measure is viewed as providing the most useful measure for making judgments about whether the association is likely to be causal, with larger relative risks more supportive of causality than smaller ones. However, for assessing public health impact, we may be more concerned about the absolute, not relative, risk of disease. We would all prefer to double our risk of a rare health problem than a common one, all other things equal.

Another measure that sometimes comes up in legal settings is the attributable fraction or proportion, which refers to the proportion of the disease among exposed individuals that results from the exposure of concern. Or equivalently, it is the proportion of disease among exposed individuals that would be eliminated if exposure were removed. When we have an individual who was exposed and developed disease, there are only two possibilities: they would have developed the disease even if they had not been exposed or they developed the disease only because they were exposed. While we cannot make the distinction with certainty, we can assess the probability that it resulted from one situation or the other. The formula for estimating how likely it was that the disease resulted from the exposure is $1-1/RR$, where "RR corresponds to the relative risk. If the relative risk is 1.0 (no effect), there is of course zero percent chance the disease was produced by the exposure. If the relative risk is 2.0 (a doubling of risk), the attributable fraction is $1 - \frac{1}{2} = \frac{1}{2} = 50\%$. That is, there is a 50/50 chance that this person would not have gotten the disease had they not been exposed. As the relative risk gets larger and larger, for example a 10-fold risk for a smoker developing lung cancer, the attributable fraction gets larger and larger ($1 - 1/10 = 9/10 = 90\%$). One of the challenges in interpreting the attributable fraction concerns the referent group, i.e., the denominator of the relative risk. Just as with the relative risk from which it is derived, the attributable risk is asking how much of the disease would be eliminated if exposure shifted from the higher level in the numerator of the relative risk to the lower level in the denominator of the relative risk. The referent group will often have "lower exposure" as opposed to "no exposure," which needs to be taken into account in its interpretation. For example, there are some environmental agents to which essentially everyone is exposed to some degree, for example, PFAS. It is simply not possible to compare "exposed" to "unexposed" because no one is unexposed, so in practice, we compare those with "more exposure" to those with "less exposure."

V. HOW EPIDEMIOLOGIC STUDIES ARE CONDUCTED

The details of how study populations are identified, how information is collected, and how data are analyzed are quite variable, but there are commonalities across all epidemiologic studies. A suitable population is chosen for the conduct of the study that includes a sufficient number of people to provide informative data on the exposures and health conditions of interest. If exposure or disease is relatively rare, this may require a very large population or choosing one that has a particularly high rate of exposure (*e.g.*, based on geographic location or occupation) or high risk of disease (*e.g.*, based on demographic characteristics). For studying the potential adverse effects of a drug, we might limit the study to people who have a condition that may require them to take that drug. For studying a hazardous chemical, we may conduct a study of those who work with the chemical as part of their job in a setting in which the chemical is frequently encountered at relatively high levels.

The goal of all scientific research is to yield generalizable knowledge, so that means the specific population we use for studies need not be the one we are ultimately interested in. We may be concerned with indoor radon and lung cancer, but more informative studies may be conducted among highly exposed uranium miners to enhance our understanding of carcinogenic effects of radon more generally. The study population is chosen to provide valid information about the cause-and-effect relationship of interest, with the knowledge that results intended for application beyond that specific set of individuals. The research is intended to provide generalizable information on whether the exposure causes disease that would apply to any population that is subject to that exposure. Valid evidence that a chemical causes disease in a worker population should be applicable to those in the community, after accounting for the different levels of exposure in the two groups. Often, risk analysis for regulatory purposes to limit air or water pollution in the community may extrapolate from occupationally exposed populations in which the evidence of a cause-and-effect relationship is clearer. The important point is that the research needs to be capable of identifying causal effects that are applicable broadly to other groups of people to be useful in addressing general causation.

Having chosen the population, there is a need to collect essential information on exposure, health outcome, and important potentially confounding factors that may also affect the risk of disease. The choice of the population may be based in part on finding groups that can provide high quality data. For example, those enrolled in longstanding health maintenance organizations may have much better documentation of their health experience than those who seek care from a wide range of providers since the former has a consolidated data resource and the latter does not.

Exposure can be measured using any of a number of specific tools but ultimately there are just a few basic approaches. People can report their own exposure through questionnaires or interviews when the exposure is known to them, such as using particular products, taking medications, or engaging in certain behaviors that might affect health such as engaging in physical exercise or working in a specific factory. Self-report can be the most accurate source of information in some cases, but does have the potential for errors in recall or misrepresentation. Sometimes the information on exposure is indirect, such as reporting where they lived with a separate step to connect that location to environmental pollutants or reporting the details of their job and from that inferring what exposures they would have encountered. Depending on the exposure pathways, the accuracy of these inferences will vary.

Alternatively, there may be archival records that can be used to determine exposure, such as databases on drug prescriptions, use of a medical device, or monitored air or water pollution. These resources often allow inclusion of very large populations since there is not a need to have contact with each individual, and they provide a systematic, objective way of assigning exposure. However, these data banks often result in some loss of accuracy when applied to each individual in the study. For example, not all prescribed drugs are actually taken and air pollution at the location of the residence may not accurately indicate levels of pollution inside the home.

A common third approach is the use of biological markers of exposure, *e.g.*, assays of blood or urine for specific chemicals. This approach has the advantage of integrating the multiple sources of exposure (*e.g.*, inhaled, absorbed, ingested, from water, air, and house dust) into a single number but can pose challenges in the willingness of participants to provide such specimens and cannot provide specific information on where the exposure actually came from. In legal matters, the concern is generally with the source of exposure (pollutant, drug, etc.). While a biomarker does indicate exposure has occurred, it is generally not possible to determine where that came from, i.e., was it in their food or drinking water or the air they breathed?

Determining health outcomes is also generally based on self-report, archival records, or clinical evaluation, with varying strengths and limitations. For conditions that are often not treated through medical care, such as dizziness or respiratory irritation, only self-report is available to assess occurrence. For more serious, treated diseases, where there is not an organized approach to health care (which applies to most of the United States), self-report of diagnosed disease may be the only or best approach since there are large numbers of health care providers and while it might be desirable is it not often feasible to access their information from every provider of care. Self-report has the potential advantages of being more complete and can include a spectrum of health conditions and information from multiple health care providers. But self-reported health experiences are inherently subjective and may be intentionally or unintentionally inaccurate.

Health care records have different strengths and limitations as a source of information on health outcomes as compared to self-report. They can include large populations efficiently and provide more objective information based on laboratory tests and a systematic approach to diagnosis, and thus are more accurate in general than self-report for many conditions. However, not all people with the same condition will seek care, making the records incomplete for those who did not, and in settings with decentralized access to care such as the U.S., the logistical issues in finding and accessing records from many health care providers may be prohibitive for large population studies even if it can be done for some individuals. There is a reason that so much epidemiologic research is conducted in Scandinavia and other settings in which health care is highly centralized and extensive records are maintained for the entire population.

A third approach to assessing health outcomes is through direct examination of individual study participants. While expensive to implement, consistent and complete information can be obtained on study participants. Collecting patient information for research allows for a wide range of measures that may fall below the threshold for clinical disease. These include clinical biomarkers like liver enzymes or thyroid hormones, or physiologic markers like blood pressure. These can be highly sensitive and measured with precision but are often only indirectly related to the more serious diseases of ultimate concern.

The other type of information that is needed concerns other risk factors for the disease of interest, potential confounders, which must be considered to isolate the potential cause of interest from other, correlated causes of disease. Often these other characteristics include sociodemographic attributes such as age, sex, and ethnicity, health behaviors such as tobacco and alcohol use, and particular exposures within the same general category as the one of interest such as other drugs or exposure to other environmental toxicants encountered in the workplace or community. The same sources are used—self-report, records, or biomarkers—with the effectiveness of those methods dependent on the details of the study population and which covariates are of interest. Demographic information may be available from self-report or medical records, tobacco use is often best assessed through interviews, environmental chemicals may require using data generated by governmental agencies, etc.

Having collected the necessary information on exposure, disease, and potential confounders, the researcher organizes the data and conducts analyses to quantify the relationship between exposure and disease, accounting for potential confounders. As noted above, the goal is to calculate and compare the likelihood of developing disease among those with greater versus lesser amounts of exposure to generate a measure of relative risk, while accounting for potential confounding factors. This entails dividing the study population into groups with differing exposure levels, which establishes the numerators and denominators for the relative risk. The numerators consist of the numbers who developed the disease within the more exposed and the denominators are the corresponding measure of disease occurrence among the less exposed groups. This information in turn allows us to calculate the risk of disease among the exposed and the risk of disease among the unexposed, which can be made into a ratio to quantify the relative risk. As a simple example, assume we are concerned with whether community residents who live proximal to a coal-fired power plant have an elevated risk of lung cancer compared to others who are not exposed to this source of pollution. We would define the geographic area considered proximal, specific census tracts around the plant, and select a comparable community with similar demographic characteristics that is not exposed. The cases of lung cancer can be identified through state cancer registries, so assume we have found 20 cases among those living in the polluted area and 18 cases in the non-polluted comparison area. Also, assume there are 6,000 residents in the area around the power plant and 8,000 in the comparison area. The overall rate of lung cancer in the polluted area would be

20/6,000 = 0.0033 or 3.3 per 1000 residents. The corresponding rate in the less polluted area would be 18/8,000 =0.0023 or 2.3 per 1000 residents. The relative risk would then be 0.0033/.0023 = 1.48. More elaborate statistical approaches are often used to account for potential confounders, but more will be said about that in subsequent chapters.

Chapter 3

INTERPRETING THE RESULTS OF EPIDEMIOLOGIC STUDIES

In this chapter we will explain the terminology and mathematical expressions used to convey the results of epidemiologic studies and how studies are interpreted. The specific sources of bias that influence measures of association are described, specifically confounding, exposure and disease measurement error, selection bias and random error. We offer guidance on how to use this framework to make an informed assessment of the meaning of research findings.

I. HOW EPIDEMIOLOGY ADDRESSES CAUSALITY

Epidemiologic studies aspire to accurately measure the causal impact of exposure on disease risk. This includes causing harm by increasing risk, having a benefit in preventing disease, or having no impact whatsoever. Recognizing that absolute certainty is unattainable, this benchmark serves as a useful tool for evaluating how close the research results are to their aspirational goal. When we use the term "cause," we mean that there are at least some people who would not have gotten the disease had they not been exposed but did get the disease because they were, in fact, exposed. The concept of counterfactuals offers a way to formalize this notion—we cannot actually observe what happens for the same individual under the condition of being exposed to the agent of concern and not exposed to the agent of concern, but that is exactly what we would like to know: For those who were exposed to the agent and got the disease, what would their health experience have been had they not been exposed? Although that cannot be answered directly, we can design studies that give an approximate answer by comparing the health experience of those who were exposed to a suitable group of people who were otherwise similar but not exposed and find out if risk of disease differs between those two groups.

Another way of describing how an exposure relates to the risk of disease considers all people to fall into one of three groups: doomed, meaning they will get the disease whether or not they are exposed; immune, meaning that they will not get the disease whether or not they are exposed; and susceptible, meaning that they will get the disease if and only if they are exposed. When we conduct studies to compare the risk of disease among those with and without exposure, we are determining how common it is to be in the "susceptible" group since only they will experience disease from having been exposed.

The validity of epidemiologic studies refers to how well the product of those studies, a measure of the association between exposure and disease, approximates the real goal of the research, a measure of the causal effect of exposure on disease. If we conduct a study and determine that people who were exposed to an environmental pollutant have twice the risk of disease as those who were not exposed, we ask whether this supports the conclusion that the pollutant has caused disease risk to double. While the cliché that "correlation does not equal causation" is true on some level, it can be countered with another cliché, "the devil is in the details." Some correlations point to causation and others do not. We must have to contend with the particulars to make an informed judgment about what the research tells us about the causal effect of interest. That requires examining the population included in the study, how exposure and disease were ascertained, what other factors that affect the disease risk may be sources of confounding, among other considerations. As explained below, there are well-established, logical, effective principals that yield an informed assessment. Subjected to this methodologic filter, experts in epidemiology can reach an informed judgment and moreover explain in non-technical terms the basis for that judgment. This does not mean that everyone looking at that same body of evidence will agree, of course, but a thoughtful, cogent explanation of the reasoning behind the final evaluation can and should be provided.

II. CAUSAL INFERENCE REQUIRES EXAMINING STUDY BIASES

The examination of epidemiologic methods requires identification of study biases that may distort the measure of association. Having identified an association between exposure and disease, there are a limited number of reasons that it may have been found—basically, it can be produced by a causal impact of exposure of disease or it can be produced by study biases. To the extent we can consider and eliminate study biases, in some cases controlling them through statistical means, what remains should be the causal impact of exposure on disease. To quote from an excellent interpreter of evidence, Sherlock Holmes, "Once you eliminate the impossible, whatever remains, no matter how improbable, must be the truth." Or to put it in more pertinent terms for epidemiology, once you have been able to discount the impact of study biases, whatever remains will approximate the causal effect.

As an example, assume we have conducted a study of processed meat consumption (hot dogs, bacon, etc. which contain nitrites) and the development of colon cancer. From that study, we have generated a measure of the relative risk of developing colon cancer among those who are heavy consumers of such foods compared to those who eat none or less of such foods. The question is whether that measure of association accurately reflects the causal effect of the foods of interest on developing colon cancer. For the purpose of assessing whether the measure of association is an accurate indicator of the causal effect (or lack thereof), It does not matter whether the study shows an adverse effect (higher risk among heavy consumers of processed meats) or not, but for this example, assume we have generated a relative risk of 2.0 suggesting a possible doubling of risk.

Now, we challenge the causal interpretation of that finding by postulating a series of sources of bias in the estimate. We might first speculate that it is simply a product of random error and that in fact, there is no effect at all. That hypothesis would be addressed by examining the confidence interval and we find that it is narrow since the study is large, say a 95% confidence interval of 1.4 to 2.8. Another speculative challenge is that those who developed colon cancer may have exaggerated their reported intake of processed meats which would create a spurious association. In response we can show that the method of assessing processed meat intake is accurate and it was assessed long before they developed the disease so that the occurrence of colon cancer could not have distorted the result. This process of hypothesizing sources of bias and examining the evidence regarding their plausibility could continue as other challenges are posed. If we are able to effectively fend off each of the challenges, at the end, we would conclude that the relative risk of 2.0 is very likely to indicate a causal effect. We cannot prove that is what it indicates, but having survived a series of challenges that would have made it a spurious positive result that is not reflective of a causal effect, we are left with more confidence that in fact it is a causal effect. We back into the inference of a causal effect in this way.

It should be noted that the need for making this interpretation applies both when a positive association is found, asking whether it reflects causal effects or study biases, and when no association is found. When there is an apparent absence of association, the question is whether that accurately reflects the absence of any causal effect of exposure on disease or whether there really is a causal effect that has been distorted and thus missed due to study biases. More generally, the question is whether the association we measure, whatever its value, accurately reflects the causal effect of exposure on disease.

To consider potential biases in a way that is helpful for inferring causality, they need to be explained in a logical way that is ultimately subject to empirical assessment. Informative data may be found within the study of concern or from other sources, but the hypothesized bias should be scrutinized for its credibility since the accuracy of the hypothesized bias bears on the interpretation of the results. There is little value in throwing out potshots without well-developed logic and ideally direct evidence, *e.g.*, "What about X?" and "Maybe there's a problem with Y." To be informative in either challenging or supporting the study's effectiveness in estimating causal effects, the potential bias needs to be considered and evaluated. The underlying concern needs to be explained clearly, with reasoning regarding what impact the source of bias would have if present, and subject to confirmation or refutation with data, i.e., it should be testable. When such hypothesized biases are put forward in this way, the evidence is examined and to the extent that relevant empirical evidence is available, the product will be a more informed assessment of causality. Either the bias will be found to be present and a source of distortion in the measure of association, making a causal effect less likely, or it is found to not be operating, in which case our confidence that the association is causal is increased for having survived this challenge.

III. SOURCES OF BIAS THAT RESULT IN INVALID RESULTS

There are different organizational schemes for potential biases, sometimes resulting in long lists of very specific concerns and laden with jargon. But for a clear understanding that is both rigorous and accessible to a more general audience, there are only a small number of reasons that an association may not reflect the causal effect of

interest. To make the assessment of biases even more feasible and informative, for a given topic and set of relevant research, there are just a few dominant concerns, typically two or three, that call for in-depth examination. Working with a short list of primary concerns tends to be more informative than an extensive list addressed superficially. For example, in studies of environmental toxicants exposure measurement and confounding are often at the top of the list of concerns. Ultimately, the epidemiology expert needs to distill the complexity arising from examining the studies through the filter of methodologic expertise and re-emerge with simple, logical explanations. The goal is to provide an accessible, persuasive rationale for the conclusions, telling the story of why each major methodologic concern arises, how the bias might operate, what evidence there is in support of or in opposition to the bias actually affecting the results, and the conclusion that can be drawn from that assessment. The menu of biases can be distilled into five pathways, with key features noted below.

A. Confounding

Confounding refers to the mixing of effects such that the exposure of interest is associated with other exposures or attributes that affect the risk of disease. When we are trying to find out if use of alcohol is related to heart disease, for example, we need to make sure that the correlation of heavy alcohol use and cigarette smoking has been accounted for since we know that smoking causes an increased risk of heart disease. If we ignore smoking, and heavy alcohol users are in fact more likely to smoke cigarettes, the association we measure and attribute to alcohol will really be distorted (increased) because of the correlation with smoking. In fact, the association is reflective of "heavy alcohol use + cigarette smoking" not "heavy alcohol use alone" and misrepresents the true effect of alcohol on the risk of heart disease. Similarly, environmental toxicants are sometimes associated with lower socioeconomic status, and there are many diseases that occur more frequently among people of lower socioeconomic status than those with greater economic means. If we fail to take this into account, what appears to be an association of pollution exposure with disease may really be a result of other correlated aspects of poverty such as poor diet or limited access to medical care.

To be helpful in interpreting results, we need to know something about what factors (other than the exposure of interest) are known to be associated with the disease and consider whether those other influences on disease are likely to be correlated with the exposure of interest. Through prior epidemiologic studies, we often have some knowledge of determinants of disease, including demographic predictors (*e.g.*, age and sex), lifestyle factors (*e.g.*, tobacco use, physical activity), and medical history. Essentially all diseases have multiple contributors, but that fact alone does not mean confounding is a problem in examining any one of those contributors since that depends on whether those other disease influences are also related to exposure. The fact that there are other causes of disease does not exonerate the one of interest. If those other influences are not associated with the exposure of interest, they do not affect the measured association between the exposure of interest and the health outcome. For example, while we all have varying genetic predispositions to disease, so long as those genetic factors are not related to the exposure of interest, *e.g.*, alcohol use, environmental pollutant exposure, then they will not interfere with our ability to study the causal effect of alcohol or pollution. They are just among the many determinants of baseline risk.

A frequent point of confusion concerns the interpretation of how multiple determinants of disease affect each other and how to interpret multiple risk factors being influential. This is distinct from confounding, in which one predictor of disease distorts the measured impact of another one. The issue here is simply that multiple factors act as risk factors for a given health condition. To use a simple example, it is known that cigarette smoking and asbestos each produce a substantial increase in the risk of cancer. Assume that smoking causes a 10-fold increased risk and asbestos exposure a 4-fold increased risk. Furthermore, we can assume that they multiply one another such that compared to a non-smoker who is not exposed to asbestos, a smoking asbestos worker has a 40-fold increased risk (4 x 10). If we are focused on whether asbestos is associated with an increased risk of cancer, we would find that it multiplies risk by a factor of 4 whether or not that individual is a smoker. In fact, given the high risk among smokers, the 4-fold increased risk due to asbestos is a much larger effect in absolute terms than it would be among non-smokers. This is counter to the notion that smoking "explains" the risk of lung cancer since it is a stronger determinant and does not leave room for asbestos to have an influence. The notion that the presence of one risk factor (often invoked for family history, for example) exonerates other risk factors (*e.g.*, environmental toxicants) is common in the lay public and often extends into legal settings, perhaps by analogy to criminal accusations where finding the culprit exonerates others. It is quite possible, in fact quite likely, that multiple factors can each contribute to the risk of disease. When we are focused on any one of those, we do need to be concerned about confounding as indicated above, for example, considering the impact exposures that are correlated with the one of interest such focusing on one drug when it is commonly used in combination with another drug to treat the same disease. But with few exceptions, we do not need to be concerned with other risk factors that are not correlated with the exposure of interest, such as genetic predisposition which affects everyone's risk to some degree but is rarely going to be associated with any exposure we encounter in the environment or consumer products. Almost all diseases are associated with age, for example, often quite strongly but we would not say that simply being old explains the occurrence of disease so effectively that their risk was unaffected by lifestyle or environmental factors. The same would be true for other determinants of risk that are not correlated with the one of concern, no matter how powerful they are.

Identifying correlates of exposure is often more challenging than identifying predictors of disease because the latter is the focus of extensive research. There are some types of exposure that have well-known correlates that may result in confounding. When studying the effects of medication use, an obvious correlate is the disease for which the medication is taken, *e.g.*, exposure to antidepressants is (obviously) associated with having depression. To the extent that the disease, rather than the drug taken for the disease, has adverse health effects, these may be mistakenly attributed to the drug being used to treat the disease, referred to as confounding by indication. Depression itself is associated with higher rates of smoking, obesity, and physical inactivity, so these characteristics will also be associated with exposure to antidepressants and need to be considered when assessing potential side effects of antidepressant use.

Unfavorable lifestyle factors are often associated with one another. Multiple conditions are associated with lower socioeconomic status, including obesity, tobacco use, and physical inactivity, for example. Environmental pollutants often go together when certain communities and parts of communities have multiple sources such as traffic-related air pollution and hazardous waste sites. Even if there is one environment of concern, a workplace for example, there may be multiple chemicals present so that isolating the health impact of anyone requires separating any effects it has from the effects of other correlated chemicals.

The wide range of exposures of interest precludes providing a generic checklist of potential confounders, but the thought process that is required is clear. There is a need to carefully examine the specific exposure of concern and determine what other factors may be related to it that are themselves associated with disease. This will sometimes be determined from epidemiologic studies that have looked at the exposure and assessed other potential correlates, but often will be found through other surveys or studies in social sciences, environmental sciences, or clinical medicine. Where there are other disease determinants associated with the exposure of interest, confounding may be present and needs to be addressed to isolate the causal impact of that exposure.

There are a number of ways that we can address confounding, so long as we are aware of its possible presence. The most definitive way is to randomly assign exposure where feasible so that we eliminate the correlation of the confounder and the exposure. When we conduct such studies, we intentionally balance the groups to be compared on demographic factors and everything else that may affect disease, such that those who are assigned the active drug on average have the same baseline risk as the group to whom they will be compared. A particular strength is that through this random assignment, we not only balance the groups on disease predictors we know about but also on those that are unknown. In effect, the only difference between the groups is the one we are interested in evaluating—the active drug versus a placebo, for example. In fact, one of the key reasons to conduct randomized trials or experiments is to isolate the exposure we are interested in from all other determinants of disease. When we do that effectively, the baseline risk of disease from the other disease determinants is equal across the groups, the ideal situation for ensuring the absence of confounding.

A recent example in which a randomized trial provided critically important information that contradicted the results of earlier observational studies was the study of replacement estrogens for women who had gone through menopause. For many years, it was believed that prescribing these hormones prevented heart disease and thus it was argued that all post-menopausal women should receive these drugs. A huge randomized clinical trial was conducted in which women at the inception of menopause were assigned replacement estrogens or placebo and they were followed for the development of heart disease. What the randomized assignment did was eliminate all the distinguishing characteristics of those who had and had not been given replacement estrogens in routine practice, which may well have been related to medical care access, having a healthy diet and body weight, getting exercise, etc., all of which could introduce confounding. By making a randomized assignment, all these factors were balanced between the drug and placebo groups so that any differences would be reflective of the effect of the hormone. In fact, there was not just a clear lack of benefit for preventing heart disease but increased risk of some other serious conditions, including breast cancer.

But for many of the exposures of interest, there are ethical or logistical constraints that make random assignment impossible. If an agent is thought to possibly cause a serious disease, for example, it would be unethical to intentionally expose people to that agent. In some cases, the exposure of interest may occur over years or decades, and even if it were ethically permissible, it would be infeasible to expect study participants to cooperate over that duration of time. Therefore, we generally need to rely on observational studies, in which the exposures were not randomly assigned but occurred due to behavioral choices, where people happen to live, the jobs that they hold, social circumstances, and other consequences of living in the real world. Intuitively, it is obvious that these exposures are far from randomly assigned and thus the potential for confounding is present. We may be interested in a specific chemical encountered in the workplace, but there are many features of the occupation other than the chemical of interest that may be related to health outcomes. The job may involve other chemical exposures, physical exertion, or risk of injury. There will be socioeconomic implications of the job, with blue collar work often related to health behaviors such as tobacco or alcohol use. A well-established influence on disease risk among industrial workers is the healthy worker effect, with active industrial workers having a more favorable health profile than the general population since such workers must have a sufficient level of physical fitness and be free of major limitations that would preclude being employed. Although we would like to study the effect of a workplace chemical in isolation, we must contend with the healthy worker effect and other disease influences that are correlated with the exposure as potential confounders.

Sometimes we have the good fortune of having an exposure that is effectively randomly allocated, not by the researcher but just happens to be independent of other disease influences. This is sometimes referred to as a "natural experiment." When the exposure occurs without awareness on the part of those who are and are not exposed, this can be effectively random. For example, drinking water characteristics vary by the water source and the distribution system serving the population, and the exposure to contaminants may differ across different suppliers. In a classic study in London by John Snow, the water supplier varied from house to house within the same area, and the risk of acquiring cholera was shown to be determined by which of two suppliers was used. When there are no structural factors that distinguish exposed from unexposed, and exposure is unrelated to economic conditions, lifestyle choices, or anything obvious, we may have the good fortune of exposure being naturally isolated from other disease determinants, and thus not subject to confounding.

More often, correlates of exposure cannot be avoided and we need to include methods for mitigating confounding. One of the ways to reduce or avoid confounding is to seek out populations in which the association between exposure and the confounding factor is absent or at least weaker. For example, whereas health care access is quite variable in populations in the United States and strongly related to socioeconomic status, this is much less of a concern in Western Europe where universal access to care is the norm. For studies in which the health outcome depends on access to and use of medical care, studies in settings in which medical care is more available will be less subject to confounding by correlates of access to care.

The choice of a comparison group to the exposed population often helps to mitigate confounding. In studying the potential side effects of a medication, for example, we may choose to compare those who took the drug due to an underlying health problem with those who did not take the drug but nonetheless had the same underlying health problem. Because we are concerned that the indication for the drug might be related to the side effect of interest (confounding by indication), we make sure that both those who used the medication and those who did not shared the underlying condition for which the drug was taken. If we are studying use of a pesticide in farming, we might choose a comparison group that is also engaged in farming but does not use that pesticide. This would effectively control confounding by the many lifestyle correlates of being a farmer.

Finally, and most commonly, we measure the potential confounding influences directly and make statistical adjustments in an effort to reduce or eliminate their impact. This requires anticipating what those potential confounders are likely to be, accurately measuring them, and then making statistical adjustments in the analysis to remove their influence on the measure of association. We must be aware of the potential for confounding based on knowledge of the influences on the risk of disease and the likelihood that those risk factors are associated with the one we are interested in. Using the example of a workplace chemical, we would ask what other influences on disease risk are associated with holding the jobs that result in exposure to that agent. We might be concerned that those who hold such jobs are prone to use tobacco or there are demands for physical fitness that make such people less likely to be obese. With this knowledge, we would want to measure and control for cigarette smoking and obesity, for example. Note that this only pertains to disease determinants likely to be associated with exposure, not to unknown genetic or other biological determinants of disease that are unrecognized and therefore not plausibly related to job choice.

The statistical methods can become quite complex and will not be presented in detail here. The key concept is that we simulate what the association between exposure and disease would have been had the groups been balanced with regard to the potential confounder. For example, assume that workers exposed to a chemical are more often smokers than the comparison group not exposed to the chemical. For example, 40% of the exposed workers smoke whereas only 10% of those not exposed to the chemical are smokers. Also assume that 10% of smokers develop the disease but only 5% of nonsmokers, i.e., a relative risk for smoking of 2.0. Even if there is in fact no effect of the chemical exposure on disease, there will appear to be an association between exposure and disease if we do not take smoking into account. The overall risk in those with the chemical exposure will be a weighted average of the 20% who smoke (with a risk of 0.10) and the 80% who do not (with a risk of 0.05), which can be calculated as (0.4 x 0.1) + (0.6 x 0.05) = 0.07 meaning that 7% of the exposed workers will develop the disease. In contrast, the risk among the unexposed is (0.1 x 0.1) + (0.9 x 0.05=0) = 0.055 meaning that 5.5% will develop the disease. The relative risk would be 0.07/0.055 = 1.27. To fix this problem and eliminate the confounding, we create two strata, one for smokers and the other for nonsmokers, and evaluate the impact of exposure within each stratum. Among smokers, the relative risk is 1.0 (no effect) and among non-smokers the relative risk is also 1.0 (no effect). When we then create a weighted average across the two strata, we get the unconfounded relative risk estimate of 1.0 (no effect).

This simple example indicates the nature of the problem – an excessive number of smokers among those with chemical exposure—and the nature of the solution, which involves stratifying by the levels of the confounder (smokers, nonsmokers) and making the comparison of exposed and unexposed within those strata. Among smokers only, there is obviously no confounding by smoking and likewise among nonsmokers. What we are doing is asking what the results would be if in fact there were equal numbers of smokers and nonsmokers in the exposed and unexposed groups. Although this was not in fact true (exposed individuals were more likely to be smokers), the statistical methods simulate what would have been found if they had been equal. For these types of statistical adjustments to be effective, we need to be aware of such potential confounders, measure them accurately, and then take them into account in assessing the association between exposure and disease.

B. Exposure Measurement Error

Accurate measurement of the exposure of interest is often challenging in epidemiologic studies. We need to start with a clear idea of what we would really like to know about the exposure that is most pertinent to the health outcome of concern. Putting aside feasibility concerns for the moment, starting with the ideal provides a benchmark to serve as a point of reference for examining what the studies were in fact able to do. For example, where chronic diseases such as cancer or cardiovascular disease are involved, we may be most interested in long-term exposure to chemicals or medications over years or decades. In practice, we can only reconstruct this history using tools such are querying people directly or looking up historical records of the jobs they held or where they lived or their drug prescriptions. When we speak about exposure measurement error, we are referring to the ways that these operational measures of exposure differ from the ideal exposure construct we are interested in.

Obviously, some exposures are easier to assess accurately than others. Use of prescription medications to treat specific diseases may be determined with some accuracy since there is mandatory documentation by the pharmacy or insurance company that provides benefits. In contrast, household use of a pesticide over extended periods of time may be hampered by the ability of people to recall what products they used or to know what chemicals are contained in those products. Additionally, in this example, whether the exposure is dermal or through breathing, the exposure may differ significantly from person to person even if they used the same pesticide depending, for instance, on whether they wore gloves and whether they were exposed in a closed environment. Similarly, biological indicators of exposure such as detection of chemicals in blood or urine may give an accurate snapshot of recent exposure and seem quite precise, but it depends on the time period of interest and how quickly the chemical is eliminated. For some chemicals that persist a long time, such as polychlorinated biphenyls (PCBs) and PFAS (per- and polyfluoroalkyl) compounds, a single measurement may reflect years of exposure, whereas for others that are metabolized quickly, like benzene and TCE (trichloroethylene), it may only reflect recent hours or days of exposure.

Beyond this intuitive assessment of accuracy based on knowing how and when the exposure occurs, we consider the data from the study or other related studies to get more direct information on how accurate the study's measurement of exposure was. In the best of studies, there may be some validation of the routine exposure measure that was used in relation to a more accurate measure of exposure, closer to the "gold standard." The more accurate measure may be too expensive or burdensome to apply to everyone, but for a subset of the study population, this measurement can be done and compared to what was used for the full study population. For example, we may want to know something about the recent use of over-the-counter medications and query people to obtain that data, but for a subset, we interview them in person and ask to see the medications they are referring to note the brand, dosage, etc. If we find that what they tell us corresponds well with what is in their medicine cabinet, we would be more confident that the information from the other participants who reported on this did so accurately.

When people have experienced a health problem, particularly a serious one, their ability to recall and report on exposures that occurred in the past may be affected and result in a more complete recall of exposure than their healthy counterparts or even result in overreporting the exposure relative to what actually occurred. This leads to recall bias, typically an overreporting of exposure relative to those who are free of the disease. This has been a concern, for example, in studies of relatively minor exposures during pregnancy such as over-the-counter drugs or use of particular household products, as they might related to having a child with a birth defect. It is easy to imagine a tendency for mothers who have had an affected child to ask themselves what might possibly have caused this condition and search their memory with more diligence than a mother who has given birth to a perfectly healthy infant. The question of what they experienced during pregnancy is much more salient to the mother whose child has had a major health problem as compared to the mother whose child is healthy. Where this sort of exposure misclassification occurs, whether it is the person free of disease underreporting actual exposures or the person with disease overreporting exposure, the consequences are predictable—it will lead to overstating the magnitude of the association between exposure and disease. The logic can be worked through in simple terms to predict the impact on the measure of association, sometimes even quantitatively if we know how often the various types of errors occur.

More generally, a critical question to ask is whether the accuracy is likely to be the same or different for those who have the health problem of interest compared to those who don't have are free of that health problem as illustrated for recall bias. This requires asking whether the health experience itself has influenced in any way the accuracy of the information we obtained about their exposure. We ask whether the exposure misclassification is differential by health status or nondifferential, meaning similarly accurate among those with and without the health problem. This type of error is always a potential concern when we obtain exposure information after the disease has already occurred. Logically, it cannot be a problem when exposure is ascertained prior to the occurrence of the disease.

When the exposure misclassification applies equally to those with and without disease, either because we assessed exposure before disease was present or without the person being aware they have the disease or when the exposure is assessed through means that are free from influence of the participant (records, laboratory tests), this results in nondifferential misclassification as described above. In this circumstance there may be errors present, discrepancies between what we would like to know and what we do know, but they apply equally across the study population. This also has predictable effects on the measure of association: relative to the true causal effect, this type of error will bias the measure of association and make it appear to be closer to the null value, understating the influence of exposure on disease. This makes sense intuitively as well—if, instead of actually measuring exposure, we flipped a coin for each person to declare them "exposed" or "not exposed," on average we would not find any association between the exposure and disease of interest since everyone was assigned exposure randomly. If only a subset of people were assigned exposure randomly, it would dilute the other subset that had accurate data and bring the group's overall results closer to finding no association. This is a common concern in epidemiologic studies that address exposures that are inherently challenging to measure for one reason or another, such as past diet or behaviors.

C. Disease Measurement Error

The issues regarding ascertainment of the health outcome, referred to as "disease" for convenience, are analogous to those for measurement of exposure, although the determinants of accuracy are quite different. In epidemiologic studies, we would like to know with certainty who has and has not developed the disease of concern, and ideally know when they first got the disease. In some cases, this is quite obvious—for example, serious, acute events like a myocardial infarction are comprehensively identified and accurately classified for the most part. But we are often interested in conditions that are far more subtle and challenging to identify comprehensively and accurately, such as development of early stages of neurodegenerative disease or subclinical changes in cholesterol or thyroid function.

Some health outcomes are only identified and documented based on the decisions made by the affected individuals to seek care that leads to a diagnosis. For example, there have been products such as the Dalkon Shield intrauterine contraceptive device (IUD) that caused pelvic inflammatory disease resulting in infertility. Not everyone who has suffered the biological event that leads to infertility would be aware of it since it would only come to attention if they attempted to conceive. Even among those who wanted to conceive and were unable to do so, some may have sought medical care to diagnose and treat the problem whereas others might not have. People may vary in their access to medical care or ability to afford the care needed to diagnose the problem, depending on their insurance coverage. For these types of reasons, patient behavior, health care access, and health care utilization, the correspondence between having the health problem of interest and being identified as having the health problem of interest is often imperfect.

There may also be challenges in making accurate diagnoses, even among those seeking medical care. Health care providers vary in their thoroughness and ultimately in their accuracy, with some health problems more easily recognized and accurately classified than others. Some psychiatric conditions, for example, can be challenging to identify accurately. A particular challenge in epidemiologic studies is pinpointing the timing of disease onset for chronic conditions. In many cases, we can determine the timing of diagnoses since that is documented in the medical record, but working backwards, it is not so clear when the patient first realized that there was a problem, and going further back, when the problem was first developing. Even for well-defined diseases like cancer, the questions of when it was developing and when it began are often different from when it was recognized.

Epidemiologic studies need to consider the way that the health condition develops, becomes recognizable, and is diagnosed and treated to judge whether a study's methods for disease ascertainment are accurate or not. The ultimate question is whether the algorithm for identifying disease is capturing all the cases that actually are present, referred to as sensitivity of ascertainment, and whether this method of identifying disease excludes those who truly do not have the disease, referred to as specificity of ascertainment. These correspond to the potential for under-ascertainment and over-ascertainment of disease. To the extent that the operational definition of disease and the truth deviate, there is misclassification, and the measure of association may not be reflective of the causal effect of

the exposure on that disease. For instance, many autoimmune diseases are extremely difficult to diagnose and even once diagnosed, there can be variation in the way such diseases are labeled by the diagnosing health care providers.

While there is always a concern when the disease assignment is inaccurate, there is a particular concern when the accuracy differs in relation to exposure. This is referred to as differential (as opposed to nondifferential) disease misclassification, that is, the sensitivity and specificity of disease identification is different for those with versus those without exposure. The impact of such errors is predictable: if those who were exposed are more likely to have their disease comprehensively identified relative to those not exposed (fewer missed cases), the measured association will overstate the impact of exposure on disease. For example, if there is a well-publicized episode of environmental contamination or concern with a consumer product, those who have been exposed to the potential harm may be more likely to seek out medical care and diagnosis than those for whom these issues are not of interest. How much impact this could have depends on how discretionary and hence incomplete the assessment of disease and how much the heightened concern affects the identification of the disease. It is also possible for those who have been exposed and have heightened concern to inaccurately be identified as having the disease even if they actually do not. But generally, the diagnosis and treatment process weeds out these "false positives" even when it allows for "false negatives" who are affected but not identified. It is important in examining epidemiologic studies not just to determine whether there may be a problem with inaccurate identification of disease but to carefully consider the process by which it would occur and the predicted impact on the overall results of the study.

D. Selection Bias from Losses to the Study Population

Because epidemiologic studies rely on human populations in the real world, it is inevitable that we will not be able to engage all those we would like to include. There will be some people who refuse to participate, cannot be located, are too sick to participate (or deceased), or are excluded for many other reasons. This loss from the study may or may not distort the measure of association. If the losses are random or effectively random, the study will be a bit smaller (and results less precise) but there will not be bias in the measure of association. However, losses are rarely entirely random.

In examining the pattern of losses, the first concern is whether it results in overstating or understating the magnitude of the relationship between exposure and disease. There are some types of losses that can be predicted to have little or no effect on the measure of association. For example, if we lose a random subset of those exposed or unexposed, and the ones that remain are essentially like those who were lost, it may be a smaller study but everything else would be accurate, particularly the rate of disease in each of those groups. Similarly, if we end up with groups that have a distinctively higher or lower risk of disease, it is not a major problem so long as that is equally true for those who are and are not exposed.

Problems arise when the losses are unbalanced with respect to exposure and disease. If the exposed people we are able to enroll in our study are especially prone to get the disease but this is not true for the unexposed people, the measure of association will be higher than it should be and overstate the impact of exposure. For example, imagine we are interested in the effect of a dietary supplement on the risk of developing heart disease and that the product was advertised as "heart healthy." When we recruit supplement users and non-users for our study, those supplement users motivated to participate may well be people who have particular concerns about heart disease because of a family history or other risk factors such as diabetes. In fact, these motivating conditions render them to be at higher risk so we will end up with supplement users who are disproportionately prone to develop heart disease. If those not using supplements are normal in all respects, not equally selected because of a higher baseline risk, the measure of association between supplement use and heart disease will be overstated relative to any true causal effect. Just as for misclassification, it is not enough to know there are losses from the study. We need to use the data and logic to figure out the pattern of those losses to predict the impact on the study's results. While this process can be somewhat complicated, like all aspects of epidemiologic reasoning, it can be explained in an accessible way with careful step-by-step reasoning.

A different problem with selection concerns whether the results for the study population are applicable more broadly to other populations of interest, referred to as generalizability. Assuming the study has accurately measured the causal effect of exposure on disease, we now ask whether the same causal effect would occur in other people—those who have a different range of exposure, live in a different area, who are of a different ethnicity or socioeconomic group, who are served by a different health care system. To the extent we have accurately identified a fundamental biological cause-and-effect relationship, we would generally expect to have it apply more broadly unless there's a specific, compelling reason not to—within reason, "people are people." But for other cause and effect relationships, there may be differences related to health care practices, culture, or psychological tendencies. The very nature of research is to seek generalizable information, but the application of findings from one population to another calls for reflection and asking the question, "Is there any reason to question whether this exposure would have the same effect in this population?"

On occasion research is done on populations that are profoundly different from the one of interest. For example, we might study pesticide exposure and health effects in a low income setting with very high exposures, high prevalence of infectious diseases, poor nutrition, and inadequate medical care. When we try to apply the results to present-day workers in a high-income setting such as the United States, the question arises as to whether the many other adverse conditions in the study setting could affect the toxicity of the pesticide of interest perhaps by making them less resilient or able to tolerate an additional stressor of this nature. A case may be made that the results are nonetheless applicable to the workers in the high-income setting, but the question is a legitimate one that calls for careful consideration.

E. Random Error

Finally, it is important to consider random variation which leads to measures of association that may deviate from the true causal effect due to chance alone. Most obvious in small studies, we may identify marked elevations in risk based on very few subjects. Likewise, a study that is very small may well fail to identify associations that are truly present. Note that this is often considered as the most important or even only concern when we test for statistical significance or conduct meta-analyses to generate pooled estimates, but this is a poor strategy for evaluating epidemiologic evidence. In later sections we will address statistical significance testing (6) and meta-analyses (Chapter 7) in more detail.

The conceptual origins of random error are somewhat abstract and mysterious—why doesn't a perfectly balanced coin that is flipped 10 times always end up with five heads and five tails? In fact, it does not of course, and sometimes, very rarely, results deviate dramatically with 10 heads or 10 tails. Putting aside all the other considerations noted previously in this chapter, measures of association generated in epidemiologic studies have this extra element of random error that can produce deviations from the true causal effect. But there are important differences between random error and the other sources of error previously described (confounding, measurement error, selection bias).

Random error gets smaller as the study gets larger unlike the other sources of bias. (Big studies are no less vulnerable to confounding or measurement error than small studies.) Just as in an experiment, the deviation between the true causal effect and the measured association shrinks as the study is expanded. In very small studies, the results may be completely unstable, with one or two changes in disease occurrence causing dramatic changes in the measure of association. One of the ways to think about how precise or stable the estimate would be to ask how many people would have to change from diseased to non-diseased or vice versa to result in a meaningful difference in the measure of association. If it's small, less than five say, that makes random error a real concern, whereas if it's large, say 20 or more, random error is not so much of a concern.

A second important feature is that with random error, small deviations from the true value are much more likely to arise than large deviations. And fortunately, small shifts are not usually a major concern in epidemiologic studies. If we measure a relative risk of 2.0, it is really not of much concern if the true value is 1.8 or 2.2, but a big concern if it's really 1.0 or 4.0. One of the ways of asking how big or small it might be is with the presentation of confidence intervals that give a sense of the range within which the true value is likely to fall. As studies get larger, that range of possible values shrinks and the range of reasonable possibilities around the measured value is narrower.

Another aspect of random error that distinguishes it from other sources of bias is that it is symmetric around the measured association, with positive and negative deviations equally likely. It is just as plausible that the association we measure is bigger than the true causal effect as it is smaller than the true causal effect, in fact by

equal amounts. Sometimes the focus is on one possible direction of random error—for example if we measure an association with a relative risk of 1.5, we might ask if there could be no effect at all, that is a relative risk of 1.0 and the measured relative risk of 1.5 is a product of random error. But it is just as reasonable to ask if there may be a much stronger effect, say a relative risk of 2.3 that was distorted (understated) as a result of random error.

In considering the array of issues that bear on the interpretation of an epidemiologic study, random error should always be considered but is often of secondary interest relative to more fundamental design issues like measurement error or confounding. It becomes a much greater concern when the health outcome is very rare, such as unusual forms of cancer or birth defects, especially when the exposure of concern is also rare. While there are formal ways of addressing random error through probability values and confidence intervals, an intuitive sense of the potential for a major effect can be inferred from the cross-tabulation of those with and without exposure divided into those with and without disease. For example, we may measure a relative risk of 2.0 in one study with a 95% confidence interval of 1.7 to 2.4 indicating a very high degree of statistical precision, whereas another study may also measure a relative risk of 2.0 with a confidence interval of 0.2 to 20.0, indicating extremely poor precision, quite possibly resulting from some groups with only one or two cases. The interpretation of those two studies that generated the same relative risk is quite different, one giving solid evidence of an association and the other completely uninformative. As a rule of thumb, when some of the cells in the calculation of the relative risk have five or fewer people in them, a reasonable rule of thumb, there is serious potential for a major impact of random error.

Chapter 4

REACHING JUDGMENTS BASED ON EPIDEMIOLOGIC EVIDENCE

In this chapter we will explain the reasoning used by epidemiologists to reach judgments about cause-and-effect relationships based on research findings. We describe the basis for interpreting associations as supporting or not supporting causal associations, and review the widely used Bradford-Hill criteria for inferring causality as well as offering an alternative approach.

I. INTRODUCTION

To assist the court and the jury, the ultimate product needed from an expert in epidemiology is an informed, carefully reasoned, evidence-based judgment about the probability that the exposure of interest acts as a cause of the health outcome of concern in the case. After scrutinizing the relevant research and evaluating the methods and results of all the important studies that bear on the question, the epidemiologist must reach a conclusion. In general, the answer will not be entirely obvious or there would not be a need for expert assessment. In fact, it seems likely that where the evidence is absolutely clear, there is less likely to be a legal dispute over causation. An epidemiology expert is not needed to dispute the evidence that space aliens cause heart disease or affirm that cigarette smoking causes lung cancer. These are settled matters and unlikely to be points of contention.

But that leaves a lot of room for territory that is subject to some degree of uncertainty. While the key question posed of epidemiology experts is often in the form of whether it is "more probable than not" that the putative cause affects the health outcome, in reality there is a continuum of evidence from 0% to 100% certainty even if it cannot be precisely or easily quantified. The above examples are fairly close to these extremes, but much more commonly the level of certainty hovers somewhere in the 20% to 80% range. This not only leaves room for disagreements in the legal setting, but there is often disagreement in the scientific community, with knowledgeable experts who are convinced that the exposure is more likely than not to be a cause of the disease and other experts who believe the evidence falls short of that threshold.

In this chapter, we will examine in more detail how epidemiologists come to conclusions from inherently incomplete information. Just as the need for one-handed economists has been argued to avoid the "on the one hand, on the other hand" advice, epidemiology experts need to work towards a bottom line judgment despite mixed evidence. Reaching an evidence-based judgment does not imply arbitrarily dismissing evidence that points in the opposite direction, but considering the totality of evidence and reaching an informed, balanced judgment. The more complex the issue and the more mixed the evidence, the greater the need for thoughtful evaluation and a clear explanation of the rationale for the conclusions that are drawn. It is in these situations where the evidence is mixed and the strength of the methodology utilized in one group of studies compared to another tips the balance that epidemiological expertise is particularly salient.

II. CONCEPT OF CAUSE IN EPIDEMIOLOGY

The way that epidemiologists examine causal effects is framed as a counterfactual statement, meaning it cannot be determined directly because it is hypothetical. The question is framed around those who have been exposed to the potential cause and developed the disease. We ask whether, if exposure had not occurred, if fewer of those individuals would have developed disease. If there are at least some such individuals, then by definition, exposure has increased the risk of disease. As noted in previous chapters, we cannot answer that question directly by rewinding the clock and having the same people relive their lives without the exposure to see what happens. That is what makes it counterfactual—we would like to compare their actual health experience from having been exposed to what their health experience would have been had they not been exposed. Design of studies, collection of data, and analysis of results are all intended to help inform the judgment about what would have happened absent exposure.

Epidemiologists address that question by considering the details of who was included in the study and how exposure and disease were measured. This information is factual, not a matter of opinion or interpretation. Likewise, the statistical results of the study are factual in nature (assuming they were not fabricated, which is

extraordinarily rare, especially in a peer-reviewed study). Measures of association are generated relating exposure to disease, which are simply statistics produced by the study. Both the methods and results can be described agnostically and even by those without deep expertise. There is no reason for disagreement at this stage. As aptly put by Daniel Patrick Moynihan, "Everyone is entitled to his own opinion but not his own facts." Expertise is needed to use that information on methods and results to evaluate whether and to what extent they support a causal effect.

Most diseases of concern have multiple contributing factors, not just a single cause. When some of the causes are known, such as demographic attributes (most diseases increase with advancing age, many are increased among persons of lower socioeconomic means), genetics (nearly all diseases have some genetic contribution), or lifestyle factors (obesity, tobacco use), that does not preclude other influences from operating. Causes of disease are not mutually exclusive from one another, and when we refer to causes we do not mean the sole cause, nor do we mean that the contributor is either necessary (the disease only occurs with exposure) or sufficient (the exposure alone causes the disease), only that the likelihood of developing the disease has increased as a result of the exposure of interest.

There may be a tendency to infer that the presence of one known cause, particularly a strong one, means the occurrence of the disease has been explained and there is no room for other exposures to contribute. For example, it is known that cigarette smokers have an increased risk of developing bladder cancer, but that does not logically mean that chemical exposures in the workplace have no impact on risk among smokers. In some cases, the presence of one risk factor (cigarette smoking) may in fact increase susceptibility to other contributors (chemicals), or they may act independently. If the chemical exposure doubles the risk of bladder cancer, it is likely to double the risk among non-smokers (who have a low baseline risk of disease) and among smokers (who have a high baseline risk of disease). In some cases, the joint effects of the two exposures may exceed what would have been expected based on each acting alone. It is rare that the two contributors cancel one another out, for example, that smokers are somehow protected from further harm related to chemical exposures. While it may be superficially logical to argue that having a known causal factor present (smoking) exonerates another putative causal factor (chemicals), that logic does not hold up under scrutiny.

When it comes to drawing conclusions about causal effects we need to recognize that the data do not "speak for themselves" but instead they provide clues that the epidemiologist examines and interprets. Epidemiologists are often quite cautious in making the leap from association to causation. There is a strong tradition of scientific conservatism, remaining skeptical and reserving judgment. While there is almost always room to acknowledge that we do not have a full understanding of the issue and that more research could shift the weight of evidence, that does not preclude making the best judgment we can using the available research at a given point in time. This is particularly so in the courtroom, where judges and juries must evaluate the existing evidence as of the time of the trial and cannot wait indefinitely for more studies to be completed to provide clarity. Expert evaluation often requires using the evidence wisely and explaining the basis for the final judgment, including recognizing ways in which the evidence is incomplete.

When experts disagree and argue about whether or not causality has been established based on the research, the debate is not about the study methods or results per se but rather about what they mean. Assessment of potential causality starts with a statistical measure of the association, often some form of a relative risk that is a ratio of the frequency of disease occurrence in an exposed (or more exposed) group divided by the frequency of disease occurrence in an unexposed (or less exposed) group. While the exact questions we ask are tailored around the specific exposure and disease, there are some generic ones that may help to explain how this works: how do we get from association to causation?

III. IS AN ASSOCIATION PRESENT?

The examination of the evidence often starts with a preliminary assessment of whether an association is present, putting aside whether any such association is causal or a product of study biases. While it is possible that a causal effect is present, but the research has failed to detect it, that would require making the inference from other lines of research despite an absence of association in epidemiologic studies. Focusing on the epidemiologic evidence alone, without some confidence that an association is present, there is no need to go further and consider whether the association is due to study biases or a true causal effect. Put simply, if there is not a greater incidence of the disease being studied in the exposed or more exposed group compared to the unexposed or less exposed control group, there is simply not support for a possible causal association and no need for further analysis.

The very presence of an association is subject to varying judgment and interpretation. Later chapters discuss statistical significance testing and approaches to combining evidence across studies using meta-analysis. While the ideal answer to the question of whether an association is present is a flat "yes" or "no," often even this simple assessment turns out to require some more nuanced interpretation of the evidence. If we were lucky enough to find a body of research that was perfectly consistent in showing a strong association or in not showing any association whatsoever, it would be easy to make the call as to whether or not it is present. But in practice, it is often somewhere in between.

Weak associations may be found that could point towards a real but small causal effect, or could be hints of a bigger effect that is blurred because of the study limitations. However, small associations could also be explained by study biases or random error. All other considerations equal, smaller associations inspire less confidence for the presence of any association being present compared to larger ones. This is one of the more compelling elements of the widely used Bradford-Hill criteria, considerations recommended for determining whether a statistical association between an environmental exposure and disease is causal (discussed later in this chapter).

Examining the pattern of results across studies is a logical way to assess the evidence for an association. All other things equal, consistently finding an association across many or all of the relevant studies inspires more confidence in an association being present than inconsistent findings in which some studies find an association and others do not. There are more subtleties here as well—there will rarely be perfect consistency across studies, but there may be a preponderance of evidence for or against an association with some outliers. A pattern of inconsistent findings can result from no causal effect with fluctuations among studies due to random error or various study biases. However, when there are notable differences in the quality of the studies (which is not uncommon in epidemiology), some may be dominant in drawing conclusions about an association, with the good studies superseding the weak studies. The overall assessment may be driven by the high-quality studies so that inconsistent findings in the poorer studies are irrelevant.

A frequent driver of study quality in studies of environmental exposures is often the quality of exposure assessment. In fact, because it is much easier to do studies with superficial exposure assessment based on residential location or job title as opposed to detailed assessment using biomarkers or environmental measurements, there is often a preponderance of weak studies and a handful of higher quality ones. This is the case, for example, with many pollutants such as PFAS or volatile organic compounds. A small number of high-quality studies or even one such study may override evidence from a series of poor-quality studies, but that assessment and interpretation calls for epidemiologic expertise and a clear explanation since those without such expertise may focus simply on counting the number of positive and negative studies. Studies of geographic variation in disease are easy to conduct using available mortality or cancer incidence data, and it may be tempting to interpret those as providing valuable information, but they rarely are effective in addressing the exposure of interest so that positive or negative findings should be viewed skeptically.

At the end of this phase of evaluation of whether an association is present, the answer to the question of whether an association is present may be yes, no, or maybe, with shades of "maybe" from "very likely" to "probably not." Unless we are confident that no association is present, the evaluation of the evidence should go forward to consider the possibility of a causal effect based on the details of the study methods. But at the end when a final judgment is rendered on whether there is a causal effect, the degree of certainty in an association being present at all is an important determinant. If we are not even certain that any association is present, the evidence for a causal effect is weaker, whereas if we are fairly certain an association is present, then something must account for that association, with the only options being a causal effect or methodologic flaws that have generated a spurious association. The cumulative assessment of the epidemiologic evidence requires consideration of both the degree of certainty that an association is present and the likelihood that the observed association is causal.

IV. DOES THE ABSENCE OF AN ASSOCIATION INDICATE THE EXPOSURE IS NOT A CAUSE OF THE DISEASE?

When the epidemiologic studies suggest that an association is absent or weak, the key question is whether we can then conclude with some confidence that the exposure is not a cause of the disease. In the legal setting, although the defense is not required to prove the negative, that is to make the case that the exposure is not a cause of the disease, in evaluating the totality of the evidence, it is often helpful to examine studies that have not found an association and ask whether these studies provide credible evidence that counters the contention that exposure is a cause of the disease.

There are several specific questions that often come up in considering whether the absence of an association provides meaningful evidence counter to a causal effect or whether the negative studies have generated spurious results, i.e., false negative findings. This becomes relevant when studies provide mixed results (which is often the case) and an assessment of the totality of the evidence requires considering how much confidence to put into those that do not find an association between exposure and disease (so-called negative studies) vis-à-vis the ones that do find an association between exposure and disease (positive studies). To the extent that the negative studies do not find evidence of an association, then the overall weight of evidence is shifted against a causal association being present.

One of the most common and credible reasons that a study may fail to identify an effect of exposure even if one is truly present is through poor measurement of exposure, health outcome, or both. In this case "poor measurement" means random error in assigning exposure or disease outcome as discussed previously. If we measure exposure by asking people about something that they are not likely to know or remember, for example, how many times they've taken aspirin in their lifetime or whether they have ever used a household pesticide with a specific active agent, the answers may be largely inaccurate. At the extreme, they are nearly random, with little correspondence between what we record and the true history of exposure. Where this is the case, the results of a study finding no association is simply uninformative with regard to a causal effect of that agent and should not be misinterpreted as providing evidence that such an association is not present. In most cases, poorly measured exposure will tend to produce spurious negative results, i.e., an absence of association, regardless of whether a causal effect is truly present. The same would be true for poorly measured health outcomes—if the assignment tends toward being random, with little correspondence between the study participants' true health status and what is recorded and analyzed in the study, we expect spurious negative results if a causal effect were actually present. In other words, under these scenarios of extensive random error in the assignment of exposure and/or disease, absence of association may mean one of two things: a causal effect is present and we have failed to identify it or no causal effect is present. Under these circumstances, the studies are uninformative and should not be given weight as evidence against an association.

The size of the study to some extent determines whether it is capable of identifying an association if one is truly present. This is sometimes referred to as statistical power—the power to identify an association of a given magnitude if it is present. The limiting factor is often the number of cases of disease, and often it is the number of exposed cases in particular, i.e., those who are both exposed and develop the health outcome of concern. Depending on the rarity of exposure and outcome, there are sometimes truly just a handful of cases, and when that number is very small, say less than five, the results are extremely unstable and may wildly overstate or understate the magnitude of association. When the addition or subtraction of one or two occurrences would drastically change the conclusions reached, skepticism is warranted. Instead of providing solid evidence for or against an association, the study may well just be uninformative regardless of the impression that is generated by the reported measure of association. Very small studies may be so plagued with random error that drawing any inferences is misleading.

Statistical significance testing is discussed in detail in a later chapter, but "non-significant" findings are sometimes interpreted as "null," but that is not correct. Null means that the relative risk is 1.0 (the incidence of the outcome being studied is equal in the exposed and control populations) or very close to 1.0, whereas "not statistically significant" allows for a range of possibilities, including indications of elevated risk that are not sufficiently precise to exclude the null value from the confidence interval. In small studies, the disparity between "not statistically significant" and "meaningful evidence against an association" is especially large. If we find a relative risk of 2.5 with a 95% confidence interval of 0.8 to 8.0, it would be incorrect to interpret that as evidence against an association. In fact, it provides moderately strong support for the presence of a positive association despite not meeting the benchmark for statistical significance. To provide meaningful evidence against an association being present, it is not enough to note that the outcome was "not significant" but rather to examine the estimated association and how precise it is. The most persuasive evidence that an association is not present would come from a relative risk close to 1.0 from a sufficiently large study with precise results and a narrow confidence interval.

V. DOES THE PRESENCE OF AN ASSOCIATION INDICATE THE EXPOSURE IS A CAUSE OF THE DISEASE?

The cliché "association does not equal causation" in fact is valid and calls for additional information to make the conceptual leap from association to causation on a case-by-case basis. Sometimes an association does provide evidence to support the inference of causation and other times it does not. Based on a full understanding of how the measure of association was generated, knowledge of the exposure and disease and how they might relate to one another, and a grasp of epidemiologic methods, an informed judgment can be provided to distinguish whether the association does or does not suggest a causal effect.

A commonly used algorithm for reaching a judgment regarding whether an association is present is the Bradford-Hill criteria (Hill, 1965),[1] which are commonly cited in legal matters. While a simpler and more direct way of asking the question of whether association indicates causation is provided below, the frequent use of the Bradford-Hill considerations calls for a clear description of the logic. As noted by the original author, the Bradford-Hill considerations are "considerations," not a checklist to be scored to reach a decision. The components serve as a reminder of some considerations that are pertinent to evaluating whether an association is causal or not, but there is a temptation to treat them as a rigid algorithm to give the impression that causation has or has not been "proven." It is commonly claimed in legal settings that causation has been proven because the criteria are met or has not been proven because one or more criteria are not met.

[1] Hill AB, *The environment and disease: association or causation?* PROC. R. SOC. MED. 1965 May; 58(5): 295–300. PMID: 14283879; PMCID: PMC1898525.

But causal inference is just that, an inference or judgment, and there is no substitute for a reasoned consideration of the evidence even if it is tempting to draw upon an algorithm like the Bradford-Hill criteria as though any objective evaluator of the evidence would have to come to the same conclusion. Ultimately the judgments of which criteria are and are not met is itself an inference and thus subject to differences of opinion.

Hill considers a series of questions to ask of the evidence after an association has been identified between an exposure and disease. Thus, this does not address the question of whether a statistical association is present, which is presumed to have been established. His lead-in question makes the purpose clear: "What aspects of that association should we especially consider before deciding that the most likely interpretation of it is causation?" This is a clear reminder that we are not being provided with a checklist that will yield some definitive judgment but rather aspects of the association to consider as we reach a judgment.

There are a number of features of the Bradford-Hill criteria that are important to keep in mind. Several of the items are intended to determine whether the observed association is due to bias since eliminating those explanations makes it more likely that exposure was a cause of the disease. Others consider evidence from lines of research other than epidemiology, such as toxicology or mechanistic studies, with the rationale that convergent evidence from multiple disciplines makes a causal effect more likely to be present. It is important to evaluate each of those lines of evidence on its own merits before integrating them into an overall evaluation—poor quality epidemiologic evidence is obviously not strengthened by evidence from other lines of research. The focus in this book is on assessing the epidemiologic evidence but the same sort of critical evaluation is needed to assess other lines of research before bringing those threads together and reaching a conclusion.

VI. BRADFORD-HILL CRITERIA FOR INFERRING THAT AN ASSOCIATION IS CAUSAL

1) Strength—When the association that is stronger, i.e., a more notably elevated relative risk, the likelihood that some confounding factor that has not been identified or addressed adequately could account for that association is reduced. A relative risk of 1.2 may well reflect confounding but a relative risk of 3.0 is less likely to be a product of an unrecognized confounder. Small associations can, of course, be causal, but to support that hypothesis we need to be confident that confounding has been fully controlled.

2) Consistency—Repeatedly observing the same association in different settings, among different populations, with different study designs, and in different time periods makes it more likely that it is causal. While unrecognized peculiarities in one study or another might produce a spurious association, it is unlikely that a series of different types of studies would all independently have some unrecognized feature that has generated false positive findings.

3) Specificity—When an exposure is associated with only one or a small number of health outcomes, it is more likely to be causal than when it is associated with many health outcomes. The reasoning is that finding an inordinate number of conditions that are associated with the exposure suggests some sort of bias. For example, when we compare reported exposures of those who have developed a disease to the reports of healthy individuals, we may find that those with the illness report a wide range of exposures more frequently than controls, suggesting a recall bias rather than causality. Of course, there are many exposures that are in fact causally associated with a wide range of health outcomes, such as tobacco smoke or ionizing radiation, but when exposures appear to be related to many different diseases, the possibility of some bias needs to be carefully considered.

4) Temporality—The exposure must precede the health outcome. This is logically true, the only absolute requirement on the list, and it would address reverse causality in which the health outcome is influencing the measure of exposure rather than the reverse.

5) Biological gradient—Observing a positive dose-response gradient in which there is a graded response to increasing amounts of exposure makes the association more likely to be causal. The reasoning here is that confounding is less likely if there is a graded response to varying levels of exposure. Potential artifactual explanations for increasing

risk across levels of exposure are less plausible than for a dichotomy of any versus no exposure.

6) Plausibility—When there is a clear biological rationale for the association being causal based on other lines of research, the presence of an association is more likely to be causal. But as noted, knowledge evolves and an inexplicable association at one point in time may become plausible at a later point in time. In addition, creative scientists can almost always come up with a theory to explain an association between exposure and disease but there is a big difference between being theoretically possible and a demonstration that it actually occurs.
7) Coherence—Interpreting the association as causal should not be in conflict with other known facts about the history and biology of the disease or be inconsistent with distinctive patterns in its occurrence. This is in a sense the converse of plausibility, with plausibility a form of affirmative evidence from other lines of research and coherence being the absence of contradiction from other lines of research.
8) Experiment—When it is possible to manipulate exposure and doing so has an effect on the health outcome, it is more likely to be causal. It may be possible to reduce exposure to determine whether in fact the risk of disease is reduced. As noted earlier, the main virtue of experiments is the ability to assign exposure randomly and thereby control for confounding by both known and unknown factors.
9) Analogy—When there are comparable exposure-disease associations that have been well-established, the likelihood that an observed association is causal is increased. If, for example, there are other similar chemicals that cause similar diseases, then when a new association between a chemical and disease is identified, a casual interpretation is favored. Also, if we know with some certainty that if an exposure causes one disease then it may be more likely that it also causes other, similar disease.

VII. ALTERNATIVE STRATEGY FOR EVALUATING CAUSALITY OF ASSOCIATIONS BETWEEN EXPOSURE AND DISEASE

While the Bradford-Hill considerations are all reasonable and may provide a helpful reminder of issues to keep in mind, they do not circumvent what is ultimately a subjective interpretation. Because they are so widely used and cited, they are often invoked to create an illusion of objectivity and certainty that is unwarranted. As an illustration of the often disingenuous claim that these allegedly objective criteria have led to an unarguable conclusion, both proponents and opponents of a causal interpretation will cite the Bradford-Hill criteria. It is invoked to suggest that they were carefully following the evidence which inevitably leads to the conclusion that they have drawn. Greater clarity may be achieved by separating the examination of the epidemiologic evidence independently to assess potential biases that could account for an established association which is what ultimately will or will not support causal inference. Subsequently, the other lines of evidence can be brought in with those lines of research also evaluated on their own merits. The judgment of how convincing and relevant toxicology or mechanistic research is, should be made by experts in those fields, not by epidemiologists who may casually invoke supporting evidence of this nature.

Focusing on the situations in which a positive statistical association is well-established, there are only a few common reasons that the positive results may be spurious. One of those is confounding—the exposure of interest is associated with other determinants of the disease and the association between exposure and disease is really due to those other determinants. This is a common problem in part because there is a tendency for bad (health-harming) factors to cluster together and for good (health-promoting) factors to cluster together. For example, assume we are interested in the health effects of neighborhood air pollution, an exposure which is only one of several ways in which economic deprivation may result in poorer health. The challenge is isolating any impact of neighborhood air pollution from other unhealthful consequences of economic deprivation such as access to good nutrition, the opportunity to engage in physical exercise, high quality medical care, etc. When we try to isolate the one we are interested in, the ability to do so successfully depends on first recognizing what those other factors are, measuring them accurately, and making the necessary statistical adjustments to remove their influence.

The analogous cluster of determinants can operate on the health-promoting side. Those who maintain a healthy lifestyle (diet, physical activity, non-smokers, not obese, etc.) tend to have lower risk of many diseases. The challenge is in isolating any one of those influences from the other when they tend to go together so tightly. An anecdotal example from a colleague noted the finding that use of sunscreen (a marker of a generally healthful lifestyle) was associated with a reduced risk of heart disease even when attempting to account for all the other factors known to affect heart disease (tobacco use, body mass index, diet, etc.). Clearly, there was not a direct causal link but in fact it was nearly impossible to remove effectively isolate use of sunscreen from the constellation of factors that define a "healthy lifestyle."

Sometimes the other determinants that act as confounders are easy to pinpoint, such as the effect of the underlying disease when trying to determine effects of a medication, or workplace chemicals correlated with the chemical of concern. With sufficient effort and attention, we can accurately identify the other factor and remove its influence from the exposures we are interested in evaluating. But when the other influences are part of a cluster related to socioeconomic background, cultural identify, a healthful lifestyle, location of residence, or type of job, it is much harder to do so effectively. Epidemiologists refer to this problem as "residual confounding," meaning even when we have tried to take care of the problem by controlling for confounding, we may well fall short because it is so difficult to fully account for the correlated factors.

Consider as an example comparing health risks associated with a particular chemical (*e.g.*, pesticide, fertilizer) used in farming by comparing farmers who use the chemical to city dwellers. We can try to account for all the ways in which farmers and city dwellers differ such as physical activity level, smoking habits, and diet, but it is highly questionable whether we can ever fully account for them. The preferable approach in many such cases is to make comparisons within the broader group that is exposed, in this case comparing farmers who use the chemical to farmers who do not. We may not be able to capture all the attributes that make those who farm unique, but if we can compare subgroups who differ in the exposure of interest but are otherwise similar, we can isolate the effect of the exposure of concern.

Another relatively common pathway for generating spurious positive results is through certain patterns of error in assessing exposure or disease. The consequences of these errors depends on their particular pattern, i.e., whether there are false positives (incorrectly assigning someone as exposed or diseased when they are not) or false negatives (incorrectly assigning someone as not exposed or free of disease when they are exposed or diseased). In addition, we need to determine whether those errors in assignment are applicable to some groups more than others, for example, if exposure tends to be overstated (false positives) more among those with disease than without. The consequences of the pattern of error need to be worked through logically to determine the overall impact on the measure of association. For example, when we tend to erroneously label people with the disease of concern as exposed relative to those without the disease, we will generate a spurious positive association, whether this results from overstating exposure in the diseased or understating exposure in those free of the disease.

To illustrate this point, if we are interviewing persons with a serious chronic disease about past exposures that are hard to recall, *e.g.*, use of over-the-counter medications, it is easy to see how their recall would be more complete or even exaggerated relative to someone who is currently healthy. Those with a serious disease tend to reflect more carefully and may well have already been ruminating to try to understand what could have caused this condition, thereby recalling exposures that a healthy person may not have remembered. For a healthy individual who never thought about such issues, recall may be incomplete. This is referred to as "recall bias" in which the greater recall among those afflicted with the disease relative to those free of the disease creates a spurious positive association. Whenever there is a fallible method for assessing exposure and an opportunity for the health outcome to affect the assignment of exposure, the study will be vulnerable to this bias.

The analogous process can apply to the erroneous assignment of disease. If there is some reason that those who have been exposed are more likely to be diagnosed or labeled as having the disease relative to those who are unexposed, a spurious positive association will be identified. This is a particular concern for health outcomes that are not fully and accurately ascertained in contrast to severe diseases that are certain to be identified. If we are studying a cause of lung cancer, a disease that is so serious that essentially anyone who develops it will be identified, we would not be concerned about selectively identifying the condition among those who are exposed relative to those not exposed. However, if the outcome is a condition like "headache" or "insomnia," there may well be a tendency for those who have been exposed to be more inclined to report such symptoms. This can occur for a variety of reasons such as greater recognition or tendency to seek health care in response to publicity about a possible effect of the exposure. In legal cases, there may be a tendency for those who have suffered harm to come forward in the hope of being compensated. This may lead to a more complete (or even exaggerated) accounting of their health problems. The stereotypical wearer of a neck brace after a minor motor vehicle collision is a cartoon version of this phenomenon.

Relative to a foolproof checklist to determine causality, this reasoning is admittedly more vulnerable to accusations of subjectivity or dishonesty to reach a desired conclusion. The key feature that allows those who are assessing the credibility of the conclusion in court or in scientific debate is the logic behind how the judgment was made. There is a need to create a clear bridge between the research and the conclusions and to do so in terms that make sense to those without training in the field. For the epidemiology expert to convey persuasive conclusions regarding the presence or absence of a causal effect, they need to have a logical argument and present it in the simplest, clearest way possible.

VIII. RELATIVE STRENGTHS AND WEAKNESSES OF STUDY DESIGNS

To simplify interpretation of an array of results when different study designs have addressed the same topic, there is sometimes an interest in legal settings of rank-ordering the quality of the evidence. While this search for a shorthand indicator of quality is understandable, particularly for those who are not deeply involved in the field, there is limited value. As indicated in Chapter 2, ecological studies that evaluated exposure and disease are of very limited value in addressing causal relationships, and case reports or case series of almost no value other than to stimulate more rigorous approaches to the topic.

As noted in Chapter 2, comparing cohort studies and case-control studies, there are common strengths and limitations, but it is more effective to examine the specific features of the study methods than to just assume that the design tells you all you need to know about the study's quality. Generalizations such as "randomized trials are good, observational studies are bad" or "cohort studies are better than case-control studies" should not be trusted as universally correct. For example, cohort studies allow for the potential of monitoring exposure and disease methodically and rigorously, whereas case-control studies do not have that capability, but rather than assuming the two types of studies follow this pattern, it is more informative to simply assess the quality of exposure ascertainment for any type of study. Similarly, cohort studies may have limited statistical power for studying rare health outcomes, but rather than assuming that is the case, it is more informative to ask directly about the size of the study and the precision in its estimates of the association.

Chapter 5

EVALUATING SUFFICIENCY OF EVIDENCE TO INFER A CAUSAL EFFECT

In this chapter we will review the methodology followed in epidemiology to judge whether or not a causal effect is really present, including the role of ancillary evidence. We enumerate arguments frequently invoked to support and arguments used to refute the inference of a causal effect based on research findings, considering their rationale and value.

I. INTRODUCTION

One of the biggest challenges for epidemiologists is to use their understanding of the research and methodologic concerns to determine whether the likelihood that the exposure causes the health effect is "more probable than not." While the tools of epidemiology are designed to determine causal effects, outside the legal setting, epidemiologists do not generally quantify their assessment or even assign adjectives in a standardized way. In fact, scientists in general are often resistant to reaching any "bottom line" conclusion because of fear that it will turn out to be wrong in hindsight, which of course is possible with incomplete information. Within the scholarly literature, there is rarely a comment that something is "80% likely" or any standardized approach to quantifying what is meant by such adjectives as "probable." In part, this is because there is no firm, objective, scientific basis for making such assignments—they are inferences that make use of the evidence but require going beyond the findings to reach a judgment. Surely if a panel of experts were convened, there would be a range of values assigned to how probable a causal association is, with no right answer, only a range of opinions, even if all the experts are working from the same evidence base and are unbiased. While making such attributions may be the norm for addressing legal issues, it is not common in the scientific arena.

Within that informal assessment of how likely a causal effect is, there is a different calculus for making errors of one type than another. Scientists are respected and rewarded for caution in making claims that do not pan out, i.e., not to overstate how strongly supportive the research is. In fact, it is perfectly acceptable, even encouraged, to remain skeptical about a causal association until it is proven with a high level of certainty. There is no shame in maintaining disbelief until convinced otherwise, whereas the culture of science looks with some disdain on those who make assertions of causal effects that are later discovered to have been in error. There are some dramatic examples of this in medicine (hydroxychloroquine to treat COVID, vaccines as a cause of autism) as well as in physics (cold fusion). Depending on how the information is being used, this may well be entirely appropriate, putting the burden of proof on those making the assertion of a causal effect and maintaining doubt until the evidence is strong enough to overcome it. But what that means is that the conventional assessment of causality relies on a threshold notably higher than "more probable than not." Instead of >50%, it may be closer to 80% or 90%. Although the statistical methods assigning probability values and determining statistical significance have at most an indirect relationship with causal attribution (discussed in detail in a later chapter), they do illustrate the level of caution in declaring results to be positive—this is not done when the probability value is 50.1% but rather 95% or sometimes even greater. Researchers are generally much more willing to fail to detect an association that is truly present than to claim an effect is present when it in fact is not.

If in fact we were able to determine the ultimate truth for a series of potential exposure-disease associations, we could compare our judgments of "more probable than not" with the truth. If we were perfect at separating them accurately based on the criterion of >50% likely versus 50% or less, and declared 100 statistical associations to be indicative of a causal effect, we would only be right for around 51 of them. If we had declared too many of them as causal, say 70 or 80, it would mean we were being too generous in making the call. Conversely if we had declared too few to be causal, say 20 or 30, it would mean we were being too conservative on average. Of course, this is not possible since we do not know the ultimate truth and make these judgments one at a time, but the concept may be useful in helping experts who normally function as cautious scientists applying a notably different threshold to apply the legal standard of more probable than not.

Applying a "more probable than not" standard also forces a dichotomization of what undoubtedly is a continuum of evidence, again something not typically done outside the legal applications of epidemiology. Using the above example of asking whether childhood vaccinations cause autism, it is not just slightly below the threshold of more probable than not, but very close to zero given numerous high-quality studies finding no association. Using the conventional dichotomy of more probable than not, there is no distinction between "essentially zero" and "some credible support but not quite at the threshold for declaring an association to be "more probable than not." Likewise, whether tobacco smoking causes lung cancer is not just barely more probable than not, it is essentially 100% certain. Again, no distinction is made between evidence that marginally exceeds 50% in the subjective judgment of the expert and a much higher probability. Simplifying results to indicate where the evidence falls in relation to the 50/50 mark is outside the norm for academic and even policy evaluations, unique to the legal setting.

Finally, it is worth noting that among a group of fully informed, objective epidemiology experts, there will be different conclusions drawn when the evidence hovers anywhere near the 50% cut-off point rather than for issues that are closer to 0% or 100%. This is not to say that either side is biased or ill-informed, only that the translation of the evidence to a conclusion is a subjective process that can lead to different judgments. To be informative in a legal or scientific arena, it needs to be more specific than just "we disagree." Identifying the points of disagreement can and should be done in a manner that is understandable to a non-technical audience, that is attorneys, judges, and juries. If the source of disagreement can be distilled into a different interpretation of a specific feature of the studies, for example, whether a confounding factor is likely to have affected the measure of association or whether the exposure was assessed in a way that would lead to an exaggeration of the association, the basis for their inference can be scrutinized and challenged. This may be the same in broad terms as the presentation of any other kind of evidence that is examined in a legal setting—the need to be able to follow the basis for assertions and inferences. Rather than simply claiming wisdom as "the expert," the underlying research and basis for interpretation of that research needs to be articulated.

II. CLARITY IN THE QUESTION OF A CAUSAL EFFECT

In addressing the question of whether a causal effect is believed to be present, there are a number of critical refinements to be considered. We can begin with the most general question which, if answered negatively, would end the deliberations: Is this exposure capable of causing disease in humans under any scenario? Stating the initial question in this manner allows for the most extreme exposure circumstances, uniquely susceptible populations, and the smallest imaginable increment in risk. We are asking, "Has or could anyone ever suffer adverse health effects as a result of exposure to this agent?" Without answering this affirmatively, all the potential refinements become irrelevant regarding exposure levels, underlying vulnerability, and magnitude of harm. If it is determined that the agent is capable of having adverse health effects, a series of refinements may then be considered to make the information applicable to the person or group of people of concern in the legal setting.

The levels of exposure in the population may vary considerably from none to very high, and it may be necessary or at least helpful to specify what the exposure circumstances are that are most applicable. For some agents, such as PFAS, there is a very low background level that all people experience, but there are subpopulations that have been exposed to a distinctly different range of exposure that is much higher. This may occur due to contamination of the water supply or due to their occupation as a firefighter who used PFAS-containing foam. When we ask if this exposure *can* cause human disease, it may be helpful to proceed from the very general answer (ever, anyone?) to specify the distinct subpopulation that is of concern (residents near a contaminated site, fire fighters). This clarification may be stated in general terms such as "those drinking water from a source with elevated levels of PFAS" to a much more specific, quantitative specification such as "those drinking water with >70 ppt PFAS" or "individuals with blood levels above 5 ng/ml." But it is worth noting that the agent may be capable, generally, of causing disease ("more likely than not") but in the population of interest, unlikely to do so.

There may also be further specification of the types of people who are potentially affected based on other determinants of susceptibility. In addressing the question of who is vulnerable to an exposure, there are three types of people though we cannot know in which group any individual belongs: those who are doomed, meaning that they would get the disease whether or not they are exposed to the potentially harmful agent; those who are immune, meaning that they will not get the disease whether or not they are exposed to the harmful agent; and those who are susceptible, meaning that they will get the disease if and only if they are exposed to the harmful agent. While this cannot be known for any individual, there are sometimes ways to better isolate the subgroup most likely to be susceptible in whom adverse effects should be more readily observed if such effects are present. For example, fetuses and infants are often more susceptible to certain types of toxicants, such as lead and mercury that affect the nervous system, as compared to older children or adults. We have known for a long time that levels of lead or mercury that might not cause a discernible health problem in adults can have clear adverse effects on infants. There is increasing interest in the elderly as a susceptible group for immunological impairment so that an agent that harms immune response to infection may not have a recognizable effect in those with sufficient reserve (younger) whereas it causes increased risk of severe disease or even death among those who had less capacity to respond (elderly). Genetic background is known to affect vulnerability to a range of diseases, and may well influence the disease burden from exogenous exposures such as pollutants or consumer products.

Epidemiology always focuses on populations and cannot pinpoint with any certainty the cause of disease in an individual. In fact, it cannot make a definitive assessment of a specific population such as residents of a given community or those working at a specific factory, which would require direct observation, contrasting the health experience of that same population but without the exposure of concern (counterfactual). However, what can sometimes be done is to narrow the population of interest to be more similar to an individual or class of concern. Any inferences about that individual or a specific community are still extrapolations from the patterns found in other populations, but the evidence from those other populations will be most relevant when the levels of exposure and other features of the population are more similar to the one of interest. If the focus of a given question is about community exposure from factory emissions, other studies of communities exposed from factory emissions will be more relevant than studies of the workers in those factories.

III. ROLE OF ANCILLARY EVIDENCE

Epidemiologic evidence is rarely the only source of information regarding a possible causal effect. Biological research is often an informative line of evidence, from purely mechanistic assessment of how exposures affect biological systems, up to and including classic toxicology, in which animals, often rodents, are exposed to the agent of concern and monitored for various health effects of concern. There may be clinical studies that include pathology of tissues thought to be affected by the exposure or other lines of biomedical research. There may be epidemiologic studies that address some intermediate biological outcomes, such as changes in blood chemistry or physiologic measures. We may be interested in the effect of a drug on the risk of heart attacks but have information on the drug in relation to blood pressure or cholesterol levels, both of which are known to increase risk of heart attacks. While such studies do not directly address the effect of the drug on heart attacks, the information on blood pressure and cholesterol contributes to the overall assessment of whether a causal effect on heart attacks is likely to be present. An overall judgment about whether a causal effect is likely to be present should draw upon the full range of such research.

While epidemiologic research provides a direct look at whether those who had higher exposures experienced greater risk of disease, there are limitations from observational studies that can be mitigated substantially but never totally eliminated. If there are multiple high-quality epidemiologic studies that all indicate an association is present and are not subject to any discernible biases, we may well conclude that a causal effect is more likely than not with little or no support from other lines of research. The issues are symmetrical for positive evidence indicating an effect as well as negative evidence indicating the absence of an effect: If there are a number of high-quality epidemiologic studies that find no association, it may safely be concluded that a causal effect is not likely to be present. The situation can become a bit more confusing when positive epidemiologic studies are accompanied by a substantial body of research that finds no support from mechanistic studies, toxicology, or clinical research. This may indicate that there is a causal effect operating in humans in ways that we cannot yet understand or the

positive results are a product of undetected methodologic errors in the epidemiologic studies. Analogously, negative epidemiologic studies accompanied by clear indications of an adverse effect from other lines of research may either mean that a plausible influence of exposure on disease is simply not occurring in humans at real-world exposure levels or that the quality of the epidemiologic studies falls short of detecting such an effect even if an effect is present. Epidemiology has a very difficult time distinguishing between "no effect" and "a very small effect" that may be too subtle to discern. Sometimes negative studies lead to the conclusion that there is no meaningful or discernible effect of exposure on disease, not to the conclusion that there is absolutely no effect.

Starting from the base of the epidemiologic studies that have been fully and properly assessed, there may be some addition or subtraction of confidence based on the ancillary evidence, but ultimately the epidemiologic research is or is not valid based on the methods used and susceptibility to bias. In assembling information in the legal setting, where these other lines of evidence are potentially important, experts in those fields should be engaged rather than counting on epidemiologists to interpret toxicology, molecular biology, or clinical medicine. While the specific considerations are different, there are studies of varying quality in all disciplines and an expert is needed to evaluate them to reach conclusions about their validity. In addition, a key question with these other lines of research is how applicable they are to the human health issues of interest. In toxicology, some species and experimental conditions are more applicable to the human situation than others, and some types of health outcomes in such studies are more directly analogous to human disease than others.

In general, these other lines of research address what *could* happen in humans who are exposed but not what *did* happen. It has been noted that for any possible exposure-disease relationship, imaginative biomedical researchers can come up with a plausible explanation for why it could happen. But there is a spectrum of "plausibility," from a very clear logical pathway from exposure to disease to a nonspecific general indication of some type of biological effect that may or may not be an indication of disease risk. There are measurable biological reactions to going from a dark room into sunlight and from being going into a hot or cold environment, but these biological responses are just indications of a physiological adjustment, not evidence that the new environment is increasing risk of disease.

There are several axes for assessing how informative the biomedical research is to the human health effects of interest:

(1) Extrapolation from high to low exposure levels, with the potential for qualitative differences when the actual mechanisms of effect are distinctive in the different exposure ranges: An example is the concern with low-level electromagnetic fields where the evidence for adverse health effects at low levels is quite uncertain (*e.g.*, cell phones) but extremely high levels of exposure clearly cause health harm (*e.g.*, microwave radiation capable of cooking).

(2) Extrapolation from the specific animal being evaluated or other biological platform (*e.g.*, cell culture) to humans: Toxicologists attempt to identify the species that most closely approximates human response or at least take that comparability into account in designing and interpreting their research. In a number of cases there are radically different responses across species, for example, PFAS is rapidly metabolized and excreted in rats whereas in humans, the exposures can persist for years, and there are large differences in PFAS metabolism by sex in rats but not in humans.

(3) Extrapolation of the experimental health indicator to the disease of concern in humans: There is a large body of experimental human research, for example, on health effects of air pollution, in which study participants are exposed to relatively low levels of such agents as ozone. This has to be to done in a manner that is certain not to result in serious or persistent health effects, but rather focuses on short-term, reversible physiologic changes. The information that is generated is quite precise but experts are needed to assess how applicable this information is to exposure over periods of years to varying, often higher levels than were used in the experiment and whether the subtle, reversible physiologic changes are relevant to the diseases of real concern.

IV. COMMONLY USED ARGUMENTS IN SUPPORT OF A JUDGMENT OF CAUSALITY

Although the specific issues differ across topics, there are some lines of evidence that are frequently drawn upon to explain why a judgment in favor of causality has been made. This list is not a checklist or algorithm that is guaranteed to lead to such inferences, but rather illustrative elements of the reasoning that is frequently used to support the conclusion that a causal effect of exposure on disease is present. Not every one of them is pertinent in every situation, but such a menu should be helpful in developing the information that will build the case for inferring a causal effect or conversely, for challenging opposing experts who argue that there is not a causal effect.

A. Statistical Evidence of an Association

The first criterion that needs to be met is evidence that a statistical association is present, a necessary but not sufficient basis for inferring a causal effect. This is often in the form of a relative risk comparing the frequency of disease among those who are exposed (or more exposed) to those who are not exposed (or less exposed). In presenting that relative risk, there is an interest both in how big it is in absolute terms and how precise it is. As you move away from a relative risk of 1.0 indicating no association, there may be a modest increase, for example, a relative risk of 1.2, a more substantial increase of 1.5–2.0, or a larger association. While it is entirely possible for a true causal effect to be small in magnitude, for example when only a subset of the population is vulnerable to the exposure, it is harder to make a convincing case for a causal effect of small associations as compared to larger ones.

An additional consideration is the precision of the estimated relative risk, often reflected in a statistical test. While this is a limited and frequently misinterpreted piece of information (*see* Chapter 6), nonetheless the claim of "statistical significance" is often invoked. For reasons discussed in a later chapter, more useful information about precision is provided by a confidence interval that reflects a range of plausible values, but with either approach the goal is to evaluate the degree of statistical support that an association is present. Evaluation of precision is an attempt to distinguish between "signal" and "noise," with small studies less able to do so with confidence, and larger studies more discerning. A small and imprecise indication of an elevated relative risk may be unpersuasive, whereas a large and precise indication of a relative risk makes the argument that an association is present more convincing. Not surprisingly, in between those extremes there is room for debate regarding the proper interpretation of the results of research. There may be large relative risks that are based on small studies and thus subject to statistical noise due to random error. And there may be modestly elevated relative risks that are quite precise, but because the magnitude of association is small, there may well be doubt about whether such evidence provides meaningful support for a causal effect. In those in-between situations, other considerations will need to be incorporated as indicated below, tipping the balance for results which are "suggestive" or "moderate" indications of an association being present.

B. Evidence of a Dose-Response Gradient

Beyond presenting the statistical results from evaluating a dichotomy of exposure (present/absent, higher/lower), there are often opportunities to study a spectrum of exposure across multiple levels (*e.g.*, none, low, medium, high). When exposure can be subdivided in this way, with more than two levels ordered from low to high, we can see whether there is a stepwise increase in risk across those levels. Our confidence in an association being present is supported when stepwise increases in exposure are associated with stepwise increases in risk of disease. If we find that the relative risk using the comparison group of "no exposure" is 1.2 for the low exposure group, 1.5 for the medium exposure group, and 2.0 for the high exposure group, this would strengthen the argument that an association is present. The potential for random error to result in the appearance of an association based on a dichotomy is considerably reduced as a cause for observing a dose-response gradient.

Even when a causal effect is present, it may not follow such a pattern if there is a threshold in which there is no effect until some critical exposure level is reached or a ceiling effect in which increasing exposure above some maximum has no further impact. In the simple case of low, medium, and high exposure, if there is a threshold that is not crossed until you reach the high exposure level, the medium exposure group will have no increase relative to the low exposure group. Conversely if there is a ceiling effect after which further exposure makes no difference, we may find that both medium and high exposure groups have equally increased risks relative to the low exposure group. However, patterns other than a dose-response gradient call for more explanation to claim support for a causal effect. An uneven pattern in which the highest risk is somewhere in the middle rather than at the highest exposure level (*e.g.*, relative risks of 1.0, 1.7, 1.3 from low to medium to high) is even less effective in arguing for the association providing support for a causal effect.

C. Quality of the Studies Finding an Association

Epidemiologic studies can vary substantially in their quality and hence vary in the confidence that can be placed in their results. Even when findings are mixed across studies, some supportive of an effect and others not, if those that are methodologically strongest tend to provide the most support for a potential causal association, the overall weight of evidence tips in that direction. Note that a selective focus on supportive studies is not cherry picking so long as the reason for placing more faith in those studies is clear. If it is based on strong methods, not just preferred results, then it is appropriate to emphasize those studies in the overall interpretation. Of course, the converse is equally true. If the methodologically stronger studies find no association with only the weaker ones indicating a possible effect, the scales tip in the other direction.

The features of the strongest studies that justify this selective focus needs to be explained in clear and simple terms to make the argument persuasive and have it be clear that there is no cherry-picking based on desired findings. For example, as noted earlier, doing a poor job of classifying exposure or disease often tends to produce null findings. Therefore, studies that use a more accurate approach to measurement and produce positive results can justifiably be cited as yielding a more accurate indication of a causal effect. In some instances, most of the studies on a topic share a weakness that only one or a small number of studies have been able to overcome. If the pattern of results indicates that the high-quality study or studies supports a causal effect, we can not only emphasize the "good" studies but explain where the "bad" studies went wrong. The explanation would simultaneously explain why one study is more informative based on the methodologic strength and why other studies are less valid based on the absence of this key feature.

D. Absence of Clear Basis for Attributing the Association to Bias

When a statistical association has been established, there are only two possible explanations for what produced it—either there is a causal effect or there is some bias in the study methods that has generated a spurious statistical association. To the extent that the potential for various biases to have produced a spurious association can be put to rest, the causal explanation is strengthened and may remain the only possible source for the association. Those who argue against a causal effect have the burden of postulating biases that would generate the statistical association, so an expert speaking in favor of a causal effect needs to examine those competing theories and counter them.

There are several different ways that candidate biases can effectively be put to rest. When the relevant studies vary in their effectiveness of avoiding the bias, yet all yield the same results, this suggests the hypothesized bias is not very influential. If a claim is made that smoking is acting as a confounder to produce a spurious positive association, yet studies that carefully control for smoking have results similar to those that do not, confounding by smoking becomes untenable as an explanation. As explained in a previous chapter, there are statistical methods of balancing smoking across groups and eliminate any effect it has and thus determine what independent effect the exposure of interest has. If in fact the hypothesized source of bias, in this case, confounding by smoking, is present, then the studies that avoid the bias can be looked to as more informative. If results from the studies that effectively control for smoking continue to show positive results, that would indicate the association is not due to confounding.

There may be ancillary evidence that addresses the plausibility of bias as a candidate explanation for positive results. For example, if those challenging a causal explanation for the association claim it is due to recall bias in which those with disease are overreporting a history of exposure, we may find evidence from other studies that this does not occur when the self-reported information is validated against objective data. Citing those methodologic studies can strengthen the credibility of research that applies those tools, showing that the postulated bias is not likely to be present.

E. Corroboration of Evidence for an Association Across Multiple Studies

A series of studies of the same topic often have differing strengths and limitations. One may do a great job measuring exposure but is vulnerable to confounding or another may have a particularly effective approach to avoiding confounding but be weaker in the approach to assessing disease. To the extent that a series of studies with varying weak and strong features all provide evidence supportive of an association, the overall case for a causal effect is strengthened. The counterargument becomes increasingly convoluted and implausible since it requires a series of assumptions regarding how these diverse sources of potential bias across studies are all leading to the same spurious association. When the constellation of evidence from across studies all point in the same direction, a causal effect that withstands various methodologic pitfalls becomes the most parsimonious and hence most likely explanation for the collective results. This assessment can be more formal in a meta-analysis, discussed in a later chapter.

F. Data on Time Trends or Geographic Patterns of Exposure or Disease

Descriptive data on patterns of disease can sometimes provide ancillary support for a causal effect. On their own, evaluation of time trends or geographic patterns is of limited value relative to high quality analytic studies since the information they generate is non-specific: many things are changing over time and geographic areas experience different disease rates for many different reasons. However, for making a judgment regarding whether positive analytic studies indicate a causal association, this aggregate data can help to provide corroborating evidence. For example, if the prevalence of the exposure of concern has undergone dramatic increases over time or is markedly higher in some areas than others, assessing whether disease rates vary in a similar way (rising over time, higher in areas with higher prevalence of exposure) complements the evidence from the analytic studies. All other things equal, variation in exposure should result in variation in disease rates, but all other things rarely are equal.

G. Biological Rationale for a Causal Effect

Epidemiologic evidence for a causal association is viewed as more credible when there is complementary evidence from toxicology or other biological support. This form of corroboration can be extremely useful since the strengths and limitations of biological research and epidemiologic studies are entirely different from one another. Laboratory research with tight experimental control provides precise information on the biological effects of the exposure of concern. Properly done, the array of concerns in epidemiologic studies are absent but of course these laboratory studies are not addressing humans or the real-world circumstances of interest. Nonetheless, they can add considerable indirect support to help make a causal inference from epidemiologic studies more compelling.

Within the spectrum of biological support, the value varies as a function of how directly relevant the information is concerning the causal relationship of interest and the species of concern (humans). Biological support may range from general evidence that the agent has measurable effects on cell cultures or animal models to a very specific indication of a clear pathway that leads to the disease of interest in a well-accepted animal model. Obviously, the more directly pertinent the evidence, the more effectively it complements the epidemiologic evidence. With well-developed animal models for cancer, for example, clear demonstration that the agent causes specific cancer types in exposed rodents can be strongly supportive of the evidence that an association found in epidemiologic studies is likely to be causal.

An example of this phenomenon can be found with PFOA exposure and testicular cancer. Early studies in rodents produced findings of what are called Leydig cell tumors, mostly benign tumors that develop from the cells in the testicles that produce testosterone. When studies done of communities in the Ohio River Valley with drinking water contaminated with PFOA showed a higher incidence of testicular cancer in the exposed populations, these prior toxicology studies strengthened the support for arguing that the observed association in humans is likely to be causal.

V. COMMONLY USED ARGUMENTS IN OPPOSITION TO A CAUSAL JUDGMENT

There is some symmetry, of course, between the arguments that can be drawn upon to argue in favor of a causal effect and those that can be used to argue against it. Failure to have the above evidence supporting causality can be viewed as ammunition to argue against causality. But the way the issues are assessed and explained does differ and includes some direct arguments against a causal effect being present.

A. Statistical Uncertainty and Cherry-Picking

While the study's results are sometimes unarguable, more often there is room for debate about how confident one can be in claims of statistical support for an association being present. In examining the evidence, an isolated measure of association with a p-value or confidence interval may be technically correct, but the broader meaning is open to interpretation. In particular, there may be reason to be concerned about whether positive findings are truly representative of the full array of study results. Variations of "cherry picking," selective emphasis on non-representative, isolated findings, are not uncommon, with researchers often hoping for positive results and digging deeply into the array of findings to highlight the ones that they "like." Even outside the legal setting, researchers often lean in this way, noting in the article abstract an isolated positive association in a sea of null findings. This is seen as a way to enhance likelihood of having the paper accepted for publication and support future work on this topic.

A thorough analysis of a rich, complex data set often generates a wide array of results. We may use different indicators of exposure, quantify exposure in different ways (*e.g.*, continuous measure, dichotomy, multiple categories), adjust for varying subsets of potential confounders, etc. A key question is what the overall array of findings suggests, not just whether there are any positive results to be found among the many that have been produced. If the data are analyzed in enough different ways and a large enough array of results are presented, it is almost inevitable that some glimmers suggestive of a positive association will be found. To quote an anonymous colleague, if you torture the data extensively enough, it will confess. Those isolated positive associations provide little or no support for the overall hypothesis that exposure causes disease. Cherry-picking in this way is appropriately open to challenge as failing to provide meaningful support for an adverse effect.

As discussed in Chapter 9, some expert witnesses have sought to base their causation opinions on a reanalysis of data presented in a peer-reviewed, published study and come to different conclusions than the authors of the published study. For the reasons set forth above, it is plausible that such opinions could be based on sound reasoning that may be more compelling than the views of the original authors, but the conclusions in the published study have been peer-reviewed and the article accepted for publication. If there were obvious flaws in the methodology or reasoning supporting the authors' conclusions, it is likely they would have been identified in the peer-review process and the study would either not have been published or edits would have been required. The "reanalysis" goes through no similar process and therefore, should be viewed more skeptically. In addition, the authors presumably were trying to explain what their results mean outside the context of an adversarial process whereas experts who re-examine the findings are doing so in support of a particular position.

B. Absence of a Dose-Response Gradient

When a causal effect is present, it is often expected that there will be a graded response to increasing levels of exposure. As noted above, there may be unevenness reflecting a threshold for any change in risk of disease or a ceiling effect when the impact on disease has been "maxed out" and increases no further. But just as a dose-response gradient supports a causal effect, the absence of such a gradient calls it into question. Whereas a dichotomy, high versus low exposure, may produce a positive association, examining multiple levels of exposure sometimes reveal an uneven and thus far less compelling pattern. For example, when intermediate exposures appear to be more strongly associated with the health outcome than high exposures, there is reason to question whether the results are supportive of a causal effect since it seems unlikely that a little bit of exposure is harmful but a lot of exposure is not.

C. Cumulative Basis for Uncertainty

While each individual source of potential bias needs to be examined for its merits, there may be a cumulative series of concerns which collectively call the strength of the evidence into question. Even if none are compelling in their own right, the overall weight of evidence is reduced to some extent by each credible source of bias that is specified, death by a thousand cuts. This is not to suggest a laundry list of generic but unsupported ways that epidemiologic studies can generate false positive findings, but rather to consider those that have sufficient plausibility and ideally have some empirical support for this particular set of studies. It is particularly relevant if the series of concerns all introduce bias in one direction, in this case towards overstating the association relative to any true causal effect. Multiple sources of uncertainty that suggest error in different directions, some exaggerating the association and others falsely reducing it, are not as persuasive that there is an overall bias towards a spurious positive association.

D. Identification and Documentation of Specific Biases that Create an Association

Hypothesized biases that lead to false positive associations can be proposed and documented empirically in some cases, subject to testing like any other type of hypothesis. This goes beyond just postulating general limitations of the relevant studies or "nitpicking," but rather a rationale for the presence of one or more specific forms of bias that can be empirically supported. In a sense, this provides evidence to explain why a statistical association was found, with the bias accounting for the association rather than a causal effect.

There are several common forms of bias that can lead to spurious positive associations: positive confounding, in which a known cause of disease is associated with the putative cause of interest (*e.g.*, smoking being correlated with caffeine intake in a study of caffeine and bladder cancer); exposure misclassification that leads to overreporting exposure or more complete reporting of exposure among those who have the disease (*e.g.*, recall bias); disease misclassification in which there is overdiagnosis or more complete diagnosis among those who are exposed compared to those who are not. These scenarios need not be complex and should be amenable to explanation in a very accessible manner that is persuasive to non-technical experts. For example, recall bias refers to overreporting of exposure among those who have or develop the disease. If the public is aware of the hypothesis that the exposure and disease are associated and reporting of exposure is subject to memory limitations or uncertainty, it may well be the case that those with the disease will more often report that they were exposed whether or not that is true. When we recall long-past medication use or pollutant exposures, imperfect recall may result in those who have given the issue a lot of attention due to their illness reporting having been exposed more than healthy individuals who have never contemplated the question before.

E. Evidence that Higher Quality Studies Are Less Likely to Identify an Association

Where there are multiple studies on a given topic, there is often a spectrum of quality which can be quite extensive, ranging from essentially uninformative studies to very rigorous, high-quality investigations. The pattern of findings across that gradient of quality is important to take note of, particularly if the weaker studies are the basis for inferring an association and the stronger studies fail to find such an association. This goes beyond saying the evidence is "mixed" or "inconsistent," and assessing whether the superior studies provide evidence supporting an effect. While the argument is less compelling, mixed findings from studies of similar quality weakens the overall case for a causal effect, but without some explanation of why this pattern occurs, it is not clear whether the studies that find an association or those that do not are more likely to be accurate.

F. Time Trends or Geographic Patterns of Exposure or Disease Inconsistent with Causal Effect

In addition to analytic studies that directly address the causal effect of interest, there may be descriptive studies of the exposure and disease over calendar time or across geographic locations. While these are generally less informative than detailed studies of individuals, they can contribute where there is a great deal of variation over time or spatially. More specifically, when the exposure to the population has increased or decreased dramatically

over time or differs markedly across geographic areas, one would expect a corresponding disease pattern to be present. As an example, there have been concerns about cell phone use being related to risk of brain cancer, yet going from no one being exposed to essentially everyone being exposed over a relatively short period of time (perhaps 20 years) has not resulted in any increase in the occurrence of brain cancer anywhere it has been studied. Unless there is some other preventive factor that has taken hold over time, this lack of an increase in disease despite a profound increase in exposure argues against there being a causal effect, independent of what studies of individuals might suggest. Studies of geographic variation such as studying rates by county or even by country are less compelling than time trends, particularly for diseases like cancer that develop over an extended period of time during which people move into and out of areas. But for an exposure that does follow sharp gradients in exposure and short-term health outcomes, failure to find that high-exposure areas have greater risk of disease than low-exposure areas may counter a hypothesized causal effect.

G. Absence of Biological Rationale for Causal Effect

As noted earlier, speculative mechanisms for disease causation can readily be identified, with rare exceptions. When there is a complete absence of a plausible biological pathway that could account for a causal effect of exposure on disease risk, the likelihood of biases as the explanation for observed positive associations is increased. In rare cases, exposure may simply be lacking in any biological response or generate a trivial, non-pathological response. Similarly, if the biological response falls within the range of what occurs in the absence of the exposure with modest fluctuations over time, the case for a causal effect on disease is weakened. This does not, of course, overcome compelling epidemiologic evidence supporting an association but when the epidemiologic evidence is mixed or marginal, absence of biological support weakens it further.

VI. CONCLUSION

For reasons indicated above, fitting the typical epidemiologic approach to assessment of causality into the legal environment has elements of fitting a square peg in a round hole. This can certainly be frustrating for potential expert witnesses and is one of the reasons that many knowledgeable epidemiologists are resistant to being drawn into legal matters. But it can also be frustrating for attorneys and for those who are the recipients of the epidemiologic information, including judges and juries. By examining the sources of that tension, we can identify general strategies for reducing it by bringing epidemiologic reasoning to bear on legal questions in a manner that is accessible and helpful. This does not mean compromising epidemiologic principles, but rather, applying them in a manner that is scientifically grounded and communicated effectively.

Some epidemiologists may find it tedious to have to explain what the research indicates, viewing it as "dumbing down" and glossing over important details because of an unsophisticated audience. However, it can be argued that the ability to put jargon aside and explain the findings and their implications in a manner that is both accurate and accessible (not one or the other) is not only possible but forces the epidemiology expert to have a deeper understanding than they would have otherwise attained. It can be edifying to dig deeply into the literature with a lens of rigorous epidemiologic methods and emerge with a clear story—what the research shows and most importantly, what it means.

Another overarching feature of legal settings that is somewhat different from the normal playing fields in which epidemiologists operate is its adversarial nature. The tension between having been engaged by one side or the other in a legal dispute and maintaining objectivity is real and can pose ethical challenges. But in one sense, the adversarial nature of these applications of epidemiology keep the experts honest. If they are not considering all the evidence in a sufficiently even-handed way, it is very likely that the other side will engage epidemiologists who can successfully challenge their assertions and prepare their lawyers for effective cross-examination.

The tools and concepts for arguing in favor of or against a causal effect should be as fully grounded in the evidence as possible and not just invoked because they help make the desired case for or against the assertion. It is a menu and not intended as a manual on how to win an argument. Even when a final judgment of whether the causal association is more probable than not have been made, there are often inconvenient countervailing lines of evidence that

need to be noted and interpreted. Defending the final judgment made about the evidence will be strengthened, not weakened by noting and considering the full set of information. That is true whether the assessment is made for scientific, policy, or legal applications.

Chapter 6

THE USE AND MISUSE OF STATISTICAL SIGNIFICANCE TESTING

In this chapter we will explain the meaning of the commonly used term "statistical significance" and the arguments put forward by those who defend the reliance on statistical tests and those who argue against the rigid application of statistical significance testing in interpreting epidemiologic data, with the latter an increasingly accepted approach to interpreting studies. The advantages of focusing on confidence intervals to characterize random error are explained, as well as the relationship of statistical tests and causal inference.

I. INTRODUCTION

The application and translation of epidemiology into the legal environment automatically creates some degree of tension between staying true to the science, which can be somewhat esoteric and obscure, and making it simple and accessible. The quote from Einstein is applicable here: "Everything should be made as simple as possible but not simpler." One aspect of the effort to achieve simplicity is the search for "bright lines" that unequivocally put the evidence on one side of the fence or another. When we ask a basic question of the research, *e.g.*, "Is there or is there not an association between exposure and disease?," the correct answer is often "maybe" or "sort of." Competing pieces of information need to be reconciled to come to an informed, defensible judgment at the end, more probable than not or less probable than not. The legal setting is not unique in this regard in that people tend to seek convenient ways to compartmentalize information even when the evidence for making the assignment falls along a continuum.

Statistical significance testing is perhaps the most common tool used for arbitrary dichotomies in research, including epidemiology, with a continuum of evidence placed decisively on one side of the fence or the other. While the final legal decision may well require a bright line yes or no distinction, the basis for making this distinction based on epidemiology or other lines of evidence is rarely so definitive. In this chapter, we consider the formal basis for statistical significance testing, how it is used in practice in epidemiology, and alternative approaches that are more helpful in addressing the goal of characterizing how precise the study results really are and whether they are supportive of an association between exposure and disease. The consumers of the information are interested in whether the results of a study can be trusted as accurate and what the magnitude of association is, should one be found. More specifically, when the study generates a measure of association that appears to be elevated, we want to know "is it really elevated or could it just be a product of random error?" Likewise, when the study generates a measure of association that appears not to be elevated, we should be asking "Is there really no association present or did the study fail to find one because the study was not large enough to find it?"

The analogy would be if we flipped a coin 10 times and got something other than five heads and five tails, say we got eight heads, we might ask, "is this a trick coin loaded to generate heads or did that just happen by chance?" To take the analogy further, we can see that getting heads six times is not much of an indication that the coin is problematic, seven times a bit more so but still not compelling, eight times and we're starting to get a bit worried, nine times even more so and 10 times is downright suspicious. With this continuum of evidence and concern, some arbitrary decision rule could be imposed—if it's eight or less, the coin is deemed to be valid, nine or more and it's not. Even in this simple example, the arbitrariness of dividing a continuum of evidence is clear and dividing epidemiologic findings into "significant" (meaning statistically significant) or "not significant" is even more misleading.

What is needed is some better way to acknowledge random error as a source of uncertainty but avoid the loss of information resulting from application of an arbitrary decision rule to determine whether the association is present or just statistical noise. Instead of making an arbitrary declaration, we might ask how strong the statistical support is for judging the coin to be faulty or for there to be an association between exposure and disease. At the end, experts need to take that into account and make a judgment, but the statistical information is just part of what they should be using and not a substitute for a careful assessment of all the relevant evidence.

To appreciate the issues involved in evaluating the statistical support for an association, we need to look more closely at the technical basis for statistical significance testing on the one hand and

examine what we really are interested in on the other hand, and reconcile the two. This does not require a deeper dive into arcane statistical theory or calculations and, in fact, requires less of a concern with what statistical theory indicates and more of a focus on what we really would like to learn from the studies. This remains an issue of some controversy within the field of epidemiology, but there has been a slow, steady recognition that statistical significance testing has been overinterpreted and inappropriately used. A recent commentary to that effect in Nature[1] offered a clear perspective on the inappropriateness of reliance on statistical testing: "We agree, and call for the entire concept of statistical significance to be abandoned." They noted that when invited to endorse their message, within a brief period 800 leading researchers from a wide range of disciplines had done so, including statisticians, epidemiologists (indeed including the first author of this book), social scientists, and biomedical researchers. To claim that statistical significance testing is widely accepted as the basis for dichotomizing study results as "positive" or "negative" is simply not true.

In legal disputes, it is natural to ask whether a particular approach to interpreting research is likely to be helpful to one side or the other. With regard to statistical testing, it can be used (or abused) by either side to bolster their case. Plaintiffs can argue that evidence of an association that falls short of attaining statistical significance nonetheless supports the claim that an adverse health effect is present. Defendants can argue that just because a particular result is statistically significant does not mean it provides meaningful support for an adverse effect of exposure on disease. Both are correct to an extent, but a more informed, nuanced interpretation would make better use of the available research for either side.

II. FORMAL BASIS FOR STATISTICAL SIGNIFICANCE TESTING

It is easiest to describe what statistical significance means where there has been random assignment, that is, exposure was assigned randomly to determine its effect on health outcome. This is familiar from drug trials where one group ends up getting the active drug

[1] Amrhein V, Greenland S, McShane B, *Scientists rise up against statistical significance*, NATURE, 2019 Mar; 567(7748): 305–307. doi: 10.1038/d41586-019-00857-9. PMID: 30894741.

and another group gets the placebo, and the decision for each individual to receive the active drug or the placebo is assigned randomly. The reason this is such a powerful study design is because random assignment will ensure that all the other factors that affect health outcomes are balanced, both those that are known (*e.g.*, age, socioeconomic status) and even those that are unknown (*e.g.*, genetic influences on disease). Given a random assignment of exposure, we can be more certain that differences are likely to be due to the drug itself. But even if we make a random assignment perfectly there will often be slight imbalances, the same way flipping a coin 10 times does not always result in five heads or five tails.

Having conducted the experiment, we now assume that there really is no effect of the treatment (the null hypothesis is true) and look at the results to see if they deviate meaningfully from what would be expected if there really is no effect, that is, equal disease occurrence in both groups. But we recognize it won't be *exactly* equal and want to see whether the deviation from a relative risk of exactly 1.00 may be just a product of random error. If we flip the coin 10 times and it comes up with something other than five heads, it may mean it is not a fair and balanced coin or it may just be random error, and we want to make a judgment of which it is. In this framework, we ask, "If the null hypothesis is really correct (no effect of treatment), how likely is this study to obtain results as or more deviant from no association as the ones that we have found?" In other words, if we repeated the study over and over, in what proportion of the trials would random error alone generate an association as large or larger than the one we have found? The rather contrived, hypothetical question is not what we are really interested in—in fact, we will not repeat the experiment over and over and all we have is the result of this one experiment. What we really want to know is how likely is it that there is an association? The temptation is to use the result of a statistical test to determine whether an association is really present but that is not what the result is addressing.

The answer to the latter question takes the form of a probability value that quantifies how likely it is that a random assignment would generate results that show as or more extreme differences between the exposed group and the controls as those that were found. This probability or p-value may range from 0.00 to 1.00, with small numbers indicating it is highly "unlikely" that the result was

obtained by chance alone. For example, let's assume the results of a clinical trial of new drug X show an improvement of symptoms in people in the group given the drug three times more often than in the control group who were given a placebo. This would result in a fairly impressive odds ratio of 3.0. But, we would also want to know how likely it is that this relative risk of 3.0 is just a statistical artifact, i.e., that there really is no benefit but the results just came out this way due to chance. If statistical analysis of this study yielded a p-value of 0.2, this would mean that if we repeated the experiment over and over, and if there were truly no effect of the drug, in 20% of the trials we would obtain indications of an association of 3.0 or greater. Not too often, but not that rare either. A p-value of 0.8 would mean that in 80% of the trials with an ineffective drug, we would find an association as big or bigger, which would lead us to conclude with more certainty that the drug really had no effect since deviations of this magnitude from the null finding of a relative risk of 1.0 would be quite frequent. If the p-value was 0.05, this would mean that repeating the experiment over and over would yield similar results of 3.0 or greater only five out of 100 times. This would provide us with much more confidence that the effect we are measuring is real and not random.

The final step to assessing statistical significance is to dichotomize the resulting p-value to guide the decision of whether we then reject or fail to reject the null hypothesis (the drug has no effect). If we fail to reject it, that is a declaration that an effect of the drug has not been proven and by default, we decide no association is present (the p-values of 0.2 and 0.8 would typically lead to not rejecting the null hypothesis because the odds of randomness explaining the result are simply too high). Conversely, if we reject the null hypothesis, we are making a judgment that there is an effect of the drug being tested. Traditionally, the magic number is 0.05 so if the p-value is <0.05 we reject the null hypothesis and if the p-value is 0.05 or greater we fail to reject (or accept) the null hypothesis. Equivalently, if we use the confidence interval (discussed below) solely to determine whether it does or does not contain the null value or a relative risk of 1.0, this is the same as testing statistical significance. As described below, confidence intervals have much greater value than simply as substitute statistical significance tests.

The reason to describe this somewhat convoluted, contrived basis for statistical significance testing is not to tout its virtues but to provide a foundation for taking findings regarding statistical significance with many grains of salt because of its limitations as a foundation for making decisions. In other words, we believe using statistical significance as a litmus test is inappropriate and results in the dismissal and rejection of meaningful scientific data that can be helpful in reaching a sound conclusion about the presence of a real association. The technology is loaded with arbitrary assumptions, convoluted reasoning, and a great deal of fiction, including infinite repetitions of the experiment. Its main virtue, perhaps its only one, requires some circular reasoning—it is useful because it's widely used and many people believe it is helpful.

Considering the rationale for generating a test of statistical significance points to several important shortcomings when it is used in epidemiology in particular:

1) In observational studies, there is no random allocation of exposure. The foundation of statistical testing is random assignment of exposure and without that, the framework for assessing statistical significance is, at best, by analogy. We are not randomly assigned to our jobs or exposure to toxicants or to medical treatment so the analogy is rather strained.
2) We do not begin with the assumption that the null hypothesis is true and that we will only be dislodged from that view with compelling evidence. In practice, we simply want to know what the causal effect of the exposure on disease accurately is, whether it is null or indicative of a harmful or beneficial effect. We are not testing a statistical hypothesis but rather trying to determine in quantitative terms how much of an effect the exposure has on disease, i.e., addressing a substantive question.
3) We are not repeating the experiment over and over to generate a probability distribution. We have results from one study, not a series of replications to tell us how unusual the results are relative to the other replications.
4) We do not need to make a firm yes/no decision based on a single study finding. Each study adds information and helps us to make a judgment. We do not need a decision rule for each finding from each study to guide us. In this way, it is, and should be, entirely different from the legal process where each case must be separately decided based only on the evidence presented in that case and not some other case. In epidemiology, the results of any particular study should be incorporated as information to help us make an integrated judgment based on a body of evidence. If we degrade the information into a declaration of "statistically significant" or "not statistically significant," important details have been lost.

Because there is every reason to believe that statistical significance testing will continue to be a prominent, sometimes dominant factor in the epidemiologic literature and the assessment of evidence by experts in legal cases, it is important to look carefully at the basis for and against reliance on this approach to categorizing evidence into separate bins definitively marked "convincing" and "not convincing". The degree to which individual experts use this statistical tool to make sharp distinctions will vary. By examining the considerations invoked by the true believers and the heretics, experts can be better able to defend their own position and counter the view of others as appropriate. Similarly, understanding the shortcoming listed above provides attorneys with tools to debunk the concept that a result from a study with a p-value of greater than 0.05 should be disregarded or that those with a p-value of 0 less than 0.05 provide definitive evidence that a meaningful association is present.

III. ARGUMENTS IN FAVOR OF RELIANCE ON STATISTICAL SIGNIFICANCE TESTING

Need for clear decision rules: Widely recognized and accepted decision rules are needed to avoid completely arbitrary, idiosyncratic, inconsistent approaches to the assessment of research findings. Without some agreed-upon framework, each expert can simply express their own view of the evidence without explaining how they came to that opinion.

Caution is needed in declaring an association to be present: Science tends to favor a cautious, conservative approach to making declarations that an association is present, with the inevitable consequence of having some false negatives (failing to declare an association to be present when it really is). The tradition is to err on the side of skepticism unless the evidence to the contrary is convincing. The framework for statistical significance testing is designed to weed out unreliable, potentially spurious indicators of an association being present and thus supports a more cautious interpretation.

Statistical testing is used widely across different disciplines: While epidemiology has its own particular strengths and limitations, and uses methods that differ from other lines of research, essentially all scientific disciplines rely on statistical testing to make judgments. This not only applies to clinical research and clinical trials in human populations but also to toxicology, social sciences, and a wide range of other fields. This commonality of approaches lends some consistency to the interpretation of evidence from multiple lines of research.

Grounded in statistical theory: While the scenario on which statistical significance testing is based is rather contrived, there is a foundation of statistical theory upon which it rests. This underpinning in a scientific, rigorous, quantitative framework is highly valued by many statisticians and gives an authoritative air to the inferences that rely on it.

Popular outside of scientific arena: Closely related to the universal use of statistical testing and the statistical foundation, the concept is familiar and widely used outside the scientific arena, including in the popular media. While there is some circularity to this, "it's useful because it's used," this may elevate its stature in providing explanations of evidence to judges and juries since they may already be familiar with the concept.

IV. ARGUMENTS AGAINST RELIANCE ON STATISTICAL SIGNIFICANCE TESTING

Arbitrary dichotomizing of study results: Putting the findings in one of two bins rather than allowing for the full range of support for an association results in a loss of information. Even if the formalities of statistical testing are accepted, making a rigid distinction between a p-value of 0.049 and 0.051 is obviously not doing justice to the data. "Barely" statistically significant is not meaningfully different from "almost" statistically significant, and yet that is the message. P-values can range from 0.00 to 1.00 and it would be more informative just to present the actual p-value that was calculated and not use categories. Even within the bins, there is a potentially meaningful difference of a p-value of 0.049 and 0.001, and likewise, a huge difference between a p-value of 0.051 and 0.999.

Random error is rarely the greatest concern in epidemiologic studies: By starting with the question "is it statistically significant" we are giving random error undue prominence in most cases. What we really want to know is how strong the study's evidence is in support of a possible *causal association* being present and focus our attention on what may be causing the measure of association that has been found to deviate from the true causal effect of interest. The most important threats to study validity should get the most attention, and those are much more often the quality of measurement of exposure or disease or confounding by correlates of the exposure of concern. In the case of very small studies, random error may in fact be a primary limitation, but we should not routinely assume that random error is paramount, which is exactly the message conveyed focusing on statistical significance.

Arbitrary decision rules create an illusion of scientific rigor: It should not be surprising that there is a desire for clear rules governing the interpretation of research, especially for experts communicating with non-experts. Unfortunately, although technical algorithms give the impression of cautious, standardized decision rules, they are not a substitute for a careful, reasoned, examination of the evidence. Substituting statistical tests for a thoughtful evaluation of the evidence may be easier for the judges and laymen to apply since it seems to generate a bright line "yes" or "no" answer to the question of whether an association is present. However, the mathematical basis for the calculation and underlying assumptions are rather esoteric and not likely to be understood by most of those who have to interpret the evidence in a legal setting. The alternative is to ask for experts to provide a carefully reasoned, clearly explained assessment of the evidence, that includes consideration of whether random error is likely to have generated a spurious association.

Statistical tests can be used deceptively in presenting data: When the focus is on each finding from the study and asking whether it is statistically significant, there is a clear opportunity for cherry-picking results. Rather than looking at the overall pattern of results, for example asking if dose-response gradients are present or if more accurate measurements produce stronger associations, the array of results may be searched to find and then emphasize those that are statistically significant or to emphasize the findings that are not. Epidemiologic studies frequently generate not one or a handful of results, but multiple tables using different analytic methods and alternative measurements. While this should be informative in making a comprehensive assessment of what the study has found, scanning for statistically significant results is a notably ineffective way to identify and describe the overall pattern of findings.

Both the value and precision of the measure of association are informative: When we describe the association between an exposure and disease, some measure such as a relative risk or odds ratio is the result. The result of that calculation, sometimes referred to as the point estimate, provides the study's best guess of the magnitude of association. In addition to that, we want to know how precise that estimate is, whether it's subject to large or small deviations due to random error. When we focus on statistical tests, we are effectively combining those two questions and losing information on either of the components. A relative risk of 12.0 with a p-value of 0.049 is statistically significant as is a relative risk of 1.2 with a p-value of 0.049. For the former, those exposed were twelve times more likely to show an effect while in the latter, only 1.2 times more likely, even though both results were deemed "statistically significant". A lot of information is lost if statistical significance is the only product of the inquiry. Similarly, both a relative risk of 1.2 and 3.7 may each have p-values of 0.051, but information is lost if both are summarized as "not statistically significant."

Study size has a major influence on statistical test results: The underlying issue in the above examples is that study size has a dominant role in determining whether results are statistically significant. Because of the nature of statistical testing, in a very small study, almost nothing will be significant and in a large enough study, almost everything will be. An interesting theoretical point is that as the study size approaches infinity, every association it generates will be statistically significant. While there are advantages to larger studies in generating more precise results, there are other important attributes of the study that need to be scrutinized to judge its value in answering the question of whether the agent being studied had a real effect. Thus, simply because a result is statistically significant does not mean it provides strong support for an association. Very small associations can be statistically significant, but unhelpful. For instance, studies in which a relative risk of 1.02 may indeed meet that threshold when based on a huge number of subjects, but we should not lose sight of the magnitude of association (a mere 2% increased

incidence) and may conclude that it is trivial or in fact largely indicating that no *meaningful* association is present at all.

Statistical testing **does not** help to judge whether finding no association provides meaningful evidence against a causal effect being present: The purpose of epidemiologic studies is to accurately estimate the causal effect, so that when we find no association, we would like to know whether that supports a judgment that no causal effect is operating. When we dichotomize results and note that a finding is "not significant" we are at best answering the question of "how strongly does it indicate an association is present?" We are not asking how confident we should be that the study has accurately indicated the absence of an association, which is an important question. If we find a relative risk of 1.0 (meaning the exposed and control groups have identical risk of disease) we want to know how precise that estimate is, just as we want to know if the result shows an elevated relative risk. Analogous to the question of whether an observed positive association may be the result of chance despite there being no real association, we would like to know if the finding of no association may be due to chance despite there being a real association present. Confidence intervals provide a tool for doing just that.

V. CONFIDENCE INTERVALS

We are all familiar with the concept underlying confidence intervals as a tool for describing uncertainty when it comes to political polling, although that term is infrequently used. If candidate X leads candidate Y 58%–42% in a poll with a margin of error of 10 points, this means the confidence interval around candidate X's support is from 48% to 68%. In other words, we are confident that candidate X's support will fall within this relatively wide range even if we are not certain about the exact value of their support. If the poll were based on a smaller sample of the population, the range around the estimate would be even wider, and if based on a larger sample, it would become narrower. Based on the hypothetical example noted above, candidate Y's support in the poll could range from 32–52%. Thus, if the polling is accurate,[2] there are plausible

[2] Our experience with inaccurate polls over the past few national election cycles lends further support to the points made above regarding the relative unimportance of random error in causing unreliable results. Those polls used

scenarios in which candidate Y could still win. If the poll showed candidate X with a 53% to 47% lead in the polls, but the margin of error was only 2%, meaning the lower limit of candidate X's confidence interval was 51%, it would be extremely unlikely for candidate Y to pull off the victory, again assuming the polling is accurate. Using this example, initially without considering the confidence intervals, candidate Y might be more discouraged by the first poll showing a 16-point deficit, but she should actually find the second significantly more troubling. The analogy to statistical significance testing would be to make declarations "candidate X will win" under some arbitrary decision rule rather than considering the estimated difference and range of uncertainty around that estimate. Moreover, the best guess is that the results will fall near the middle of the confidence interval, with the values at the boundary plausible but not the most likely to occur.

Confidence intervals can be constructed around the estimate of the relative risk to describe the range of uncertainty. The traditional basis is to construct a 95% confidence interval, such it will contain the correct value 95% of the time. This can be used as a substitute statistical test in that if any part of the range falls below 1.0 in a 95% confidence interval, then the result would not be "statistically significant." However, properly used and interpreted, confidence intervals provide much more information than a statistical test. They convey a clear sense of what the value is most likely to be, with the point estimate itself most probable but values around that in both directions quite plausible and those more distant increasingly unlikely. It also shows the difference between large studies that generate precise results and small studies that generate imprecise results. The range in which the true value probably lies is the focus, as it should be.

A relative risk of 1.0 with a 95% confidence interval of 0.9–1.1 would not be statistically significant, but moreover suggests that an increased or decreased risk is unlikely to be present. It provides precise evidence against an association being present. In contrast, a relative risk of 1.0 with a confidence interval of 0.1–10.0, also "not significant", provides little evidence that an association is not present since the results are so imprecise. To summarize both as "not statistically significant" and provide no other information fails to consider the certainty that an association is not present.

The same reasoning applies to positive results. If one study finds a relative risk of 1.2 with a confidence interval of 1.1 to 1.3, and another finds a relative risk of 5.0 with a confidence interval of 3.2 to 7.5, a great deal of information is lost by simply noting both are "statistically significant." The former suggests there is likely to be a very small association and the latter indicates there is likely to be a large association present.

By presenting estimates of relative risk with confidence intervals, not just statistical test results, we gain a great deal of information to answer the question that motivates the use of statistics in the first place—what is the best estimate of the size of the association and what is the range of uncertainty around that estimate? Both are important in judging whether there is a relationship between exposure and disease and in predicting the election results from political polling.

VI. STATISTICAL SIGNIFICANCE AND CAUSAL EFFECTS

The underlying purpose in evaluating the results of epidemiologic studies is to help inform a judgment regarding whether there is a causal effect. Statistical significance testing is often used as the first step in considering the possibility of a causal effect, namely addressing the question of whether an association is present, putting aside the question of whether that association is causal. The use of statistical significance testing provides a seemingly clear answer—yes or no, an association has been found or an association has not been found. It seems that this desire for clarity is what makes this testing so appealing. If the association found in the research is not statistically significant, the interpretation is that no causal effect is present. If the association is found to be statistically significant, then we would proceed to consider whether that association is causal. This is not just the case in the legal applications of epidemiology but common in the practice of epidemiology more generally. Perhaps reflecting the desire for positive results on the part of researchers and those who are seeking actionable findings to advance public health, there is a tendency to give more attention to positive results. For example, the Bradford-Hill considerations are predicated on an

the same type of statistical testing to predict the range of random error. But the results fell outside that predicted range, indicating some other flaws in the polling besides random error. Similarly in large epidemiology studies, selection bias or exposure measurement errors are much more likely to produce an inaccurate result than random error.

association being present and help to inform a judgment as to whether that association reflects a causal effect.[3]

The conventional approach is to challenge positive findings of an association by asking whether it is causal or a result of bias, but the exact same questions should apply to finding no association—whether it accurately reflects an absence of a causal effect or is a result of bias.

While it is important to make some assessment of whether an association is present, for reasons discussed above, testing statistical significance does not deliver the definitive answer that it may appear to provide. Rather than isolating the consideration of random error from other factors that bear on a causal interpretation or treating random error as the most important or first consideration in assessing causality, it should be viewed as one of several reasons that the association we measure may not accurately reflect a causal effect. The underlying question that should be asked regarding the study results is whether they accurately reflect the causal effect of exposure on disease or whether they are distorted for some reason. One of those the possible reasons they may be distorted is random error. In a small study, it is certainly possible to observe a positive association that in fact just happened to be an aberration, analogous to the normal coin that turned up 10 heads in a row. But rather than making a formal declaration at this point, yes or no, it is more informative to describe what the data show and how plausible it is that the observed association is a product of random error alone. A more accurate characterization acknowledges the continuum of certainty regarding whether an association is present as part of a comprehensive evaluation of what the study or studies are telling us. Confidence intervals provide a more informative basis for making this judgment.

Relative to this tempered interpretation of statistical significance, the inferences drawn from statistical significance testing tend to be too extreme in one direction or the other. If a result is "not statistically significant," that does not mean that a null association was found, that the relative risk is 1.0, or that a causal effect has been disproven. The measure of association may be markedly elevated or reduced, quite deviant from 1.0, (for example, it may be 3 or 4) but just not meet the less litmus test of having a p-value of 0.05 or less. Thus, a finding of "not statistically significant"

[3] *See* Chapter 4 for a discussion of the Bradford-Hill criteria for evaluating causation.

should not be viewed as exonerating the exposure of concern. Likewise, finding an association that is statistically significant does not necessarily mean there is likely to be a causal effect. In a very large study, tiny elevations in risk may be statistically significant but too small to be worthy of consideration. In very large datasets, it is not unusual to see relative risks of 1.05 as statistically significant but that may be so close to 1.00 as to be uninformative regarding a causal effect. And of course, even a more substantial association may be the product of confounding, measurement error, or selection bias, which would often be statistically significant but not supportive of a causal effect. It has sometimes been noted that large studies generate precise results, but if the methods are flawed, the results are precisely wrong.

For these reasons, it is more informative to relegate random error to just one of several reasons that the measured association may not accurately correspond to the causal effect. In some studies (small ones) it can be a very important concern and in other studies (large ones) a very minor, even negligible concern. Just as we consider other sources of error and assess how likely each of them is to be present and how much distortion each may have introduced, we should do the same with random error. But just as for other sources of uncertainty, the assessment of random error requires thoughtful evaluation, including careful evaluation of the precision of the measured association, most effectively by examining the confidence interval. Such a careful assessment produces an informed, nuanced judgment of how much of a concern random error really is.

VII. RECOMMENDED APPROACH TO ASSESSING THE IMPACT OF RANDOM ERROR

A concern with random error is justified since the exact measurement of the association may well deviate from the causal effect simply for this reason. It is one of the sources of uncertainty in epidemiologic study results and should be addressed. But as discussed earlier, we need to begin the assessment with a question of what the most important, influential, and plausible sources of bias in the studies are likely to be. If the study is very small, random error may be a major, even dominant concern, but as studies get larger and larger, the concern diminishes relative to other issues. We should not assume at the outset that random error is the dominant issue and consider it before all other issues and where a p-value falls slightly above 0.05 discard the results entirely or worse, assume that this means an association has been disproven.

What is needed is a way to quantify some basic ideas, namely that big studies are less affected by random error than small studies, and that random error is more likely to generate small deviations from the point estimate of the association than large deviations. The use of confidence intervals, appropriately interpreted, can provide that information. The traditional way this is done is by defining a 95% confidence interval around the estimate of the association, which provides a range of values with the correct one (which is unknown) likely to be contained within that range. As studies get larger, the width of that interval becomes narrower, a way of indicating that extreme deviations from what was found are increasingly unlikely. It is useful to think of the interval as describing the classic bell curve of probabilities, with the best guess being the value for the association that was measured, values around it nearly as likely to be true, and a diminishing likelihood as you move towards the upper and lower boundaries of the interval.

When we have that information, an estimate of the size of the association and confidence interval, we can make an informed assessment of how strong the evidence is for an association being present. There is no need to dichotomize results and make a declaration of "present" or "absent," but rather to consider more fully what the results tell us. For example, if there is a small association identified but it's very "noisy" because it came from a small study, as indicated by the wide confidence interval, we might legitimately relegate the research as not being supportive of an effect. In contrast, a small association based on a large study with a narrow confidence interval may be interpreted as providing meaningful evidence of a small effect being present. Likewise, even a large association from a very small study may not be very persuasive, but proper interpretation accounts for both the large association and the substantial statistical uncertainty associated with it, not just one or the other. With an estimated association and confidence interval, we can separate the "best guess" from the statistical uncertainty in the best guess, and not lump those issues together as statistical significance does.

Despite the many nuances of statistical testing and the questionable value for interpreting research, it is important to have a way to take random error into account in interpreting studies.

Abandoning, or at least downplaying statistical significance testing does not lead to anarchy in the form of a free-for-all of interpretation. Just as informed epidemiologists can examine measurement error or confounding in a thoughtful, defensible manner, they should be treating random error in a comparably sophisticated, evidence-based manner.

Judges and juries can easily become overwhelmed when trying to understand epidemiological concepts and evaluate the strengths and weaknesses of competing epidemiological studies and epidemiology expert opinions. Applying statistical significance testing as a litmus test may be enticing as a simplistic approach allowing them to arrive at the definitive result required in court. It is important for experts and attorneys to educate judges and jurors to recognize that such an approach, while easier, is far from the best method for them to fairly decide whether a causal association is present.

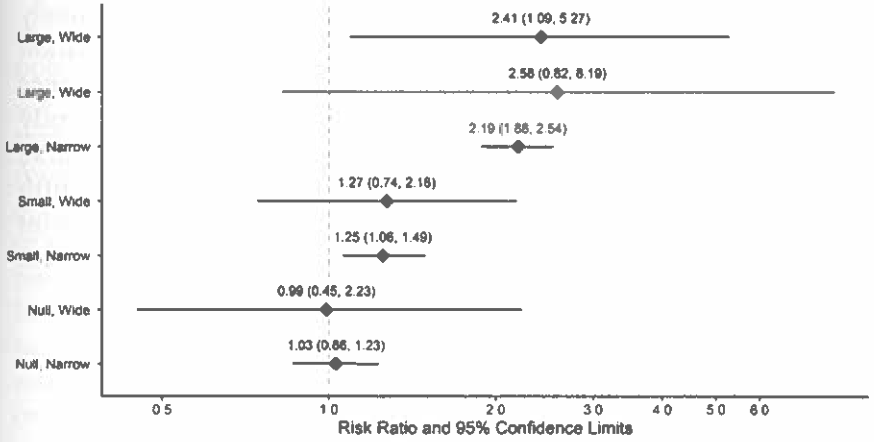

Figure 1. Graphical Illustration of Confidence Intervals

The key points in this chapter are illustrated in Figure 1 above, which shows relative risk estimates with a range of point estimates, from strong positive association to weak positive association to no association, and also has examples with wide and narrow confidence intervals, some of which cross the boundary of 1.0 (and are thus not statistically significant) whereas others do not (and are statistically significant). There are several key points to note from this figure:

1) Statistical significance is not an effective basis for characterizing the results of a study. The first, third, and

fifth examples are all statistically significant, but have notably different implications regarding the presence and strength of an association. Likewise, the second, fourth, and sixth relative risks are not statistically significant, yet the second shows a rather strong association, the fourth a weak, perhaps negligible association, and the sixth and seventh show no association with the seventh providing strong evidence against an association.

2) Precision is always an important consideration, whether the relative risk is substantially increased, modestly increased, or null, illustrated in the seven relative risks depicted in the figure. It provides an indication how much noise or random error may be affecting the relative risk, which provides informative evidence on how much confidence to have in the point estimate of the relative risk.

3) The point estimate of the relative risk is always an important consideration since it constitutes the most likely value. Considering that point estimate in combination with information on its precision based on the confidence interval provides an accurate sense of what the study results show. Whether the association is sizable (first, second, and third examples), modest (fourth and fifth examples) or null (sixth and seventh examples), the point estimate provides an anchor to be combined with information on precision.

Chapter 7

INTEGRATING EVIDENCE ACROSS STUDIES

In this chapter we will discuss the rationale for integrating evidence using meta-analysis to assess the presence or absence of an association and the limitations of this approach for addressing causal effects. An alternative approach groups studies based on their methods to assess the impact on the pattern of results. The application of evidence synthesis in legal settings is also considered.

I. INTRODUCTION

As discussed in previous chapters, epidemiologists are rarely comfortable or on solid ground relying on a single study to evaluate potential causal effects. When there are multiple studies addressing the same question, the judgment regarding whether a causal effect is present obviously should make use of the full array of relevant research. The availability of multiple studies allows for examination of consistency across studies in the replication of findings, comparison of results from studies that used different methods, and pooling of results across studies to reduce random error. While the potential benefit of identifying and considering all relevant studies is unarguable, the rationale and methods by which the constellation of research results are considered is often a complex and sometimes controversial process.

In epidemiology, the concept of pure replication is not applicable. In the laboratory, an experiment can truly be repeated using the same biological system (*e.g.*, a particular genetic strain of mouse) and the same exact exposure (*e.g.*, agent, dosage, timing and mode of administration). In epidemiology, the specific population being studied is always different and in general, the other features of the study such as the exposure and how it is measured, the health outcome and how it is measured, and how the data are analyzed and interpreted will all differ to at least some extent. As discussed below, we can pretend that none of those differences are present or that they are unimportant and lump the findings together across studies, treating a series of studies in effect as one big study. Or we can look more closely at each study's features and draw insights from the differences that help in interpreting the results. Most importantly, if we have a set of studies that use different methods, some better than others, there is the ability to see how the character and quality of the methods affect the results. If the studies had all done exactly the same thing (or we pretend that is what was done), the opportunity to assess how the methods relate to the results is lost and we miss the opportunity to obtain a more informed, persuasive understanding of what the body of research means.

Just as when we evaluate individual studies, it is important to keep the key questions in mind and make sure the approach to integrating evidence is being optimally used to help us answer those questions. The purpose of looking at specific studies or the collection of studies is to determine whether the measures of association they produce can be interpreted as measures of the causal effect of exposure on disease. Therefore, when we consider how to make optimal use of an array of studies, we are focusing on how to use their similarities and differences and the constellation of study results to help make that judgment: is the literature indicative of a causal effect?

II. RATIONALE FOR CONDUCTING META-ANALYSES

An increasingly popular approach to examining a set of studies addressing the same question is through a systematic review that generally culminates in what is referred to as a *meta-analysis*. The product of a meta-analysis is often a pooled estimate of the association between exposure and disease, combining the results from a series of studies into a single number, the pooled relative risk or odds ratio. The intention, of course, is that this pooled estimate is the most accurate indicator of the effect of exposure on disease risk because it is based not on the data from a single study, but on the combined data from a number of different studies. The reported association is a weighted average across the studies, with big studies contributing more and small studies contributing less. Since it draws on information from many studies, the estimate is often very precise with a narrow confidence interval. The approach has been used for a long time to synthesize results from clinical trials. Meta-analyses can be particularly useful when all the trials used essentially the same protocol but, the individual studies were small and each was therefore limited in regard to the precision of the results. Under those

assumptions: essentially identical protocols and random error as the primary source of uncertainty, a pooled estimate is quite useful as the best estimate of the effect of the intervention on health outcomes.

There are a number of positive features but also some serious limitations in systematic reviews and meta-analysis. On the positive side, the review often begins with a methodical approach to identifying and screening all the literature that may contribute to helping answer the question of interest. The methods used for searching databases for relevant studies and finding every possible article ensures that those conducting the review are not cheating by only including studies that they "like" because of their results or other features or simply that they have failed to take advantage of all relevant research. It provides an objective approach to identifying all the studies that have addressed the question of interest.

The next step often involves screening a large number of possibly contributory studies, again using a well-defined protocol to winnow down the list to the ones that will be most informative. Results are then extracted from each of the studies to generate comparable information that can be pooled. All of the steps from defining the question to generating results from relevant studies are commendable because they are done methodically and explicitly described.

By pooling the results across studies, the fluctuation in findings due to random error is minimized, which is why the confidence interval becomes narrower. The weighted average gives large studies more credit and small studies less credit, smoothing out differences that would otherwise be distracting from the overall pattern. The random quirks in individual studies in which some find aberrantly strong or weak associations are smoothed out when we combine the results across all the studies. If we are trying to estimate what proportion of times flipping a coin will generate heads and tails, rather than being distracted by the trials in which we obtain 80% heads or 10% heads, we find across a series of trials that the best estimate approaches 50% heads.

III. DISADVANTAGES OF RELYING ON META-ANALYSIS TO SYNTHESIZE EVIDENCE

The goal that meta-analysis addresses can be described more generally as "evidence synthesis," emphasizing that the array of informative research, the evidence, is brought together, synthesized, to make an informed judgment regarding causal effects. Generating a pooled estimate is one tool that can be considered. It is by no means the only way to synthesize evidence from an array of studies, nor is it necessarily the optimal way to do so, especially for epidemiologic studies. Examined through that wider lens of evidence synthesis, there are features of meta-analysis that often result in failure to take advantage of the full array of information and fall short of considering the important methodologic issues and differences in the studies. Looking to meta-analyses as the default approach to combining evidence across studies has, unfortunately, become routine in epidemiology and is often oversold as the only way or the most rigorous, objective way to interpret the evidence. The findings from meta-analysis may in fact overstate or understate the extent to which a set of studies support a causal effect of exposure on risk of disease.

The first steps in a systematic review, identifying all potentially relevant studies and screening the studies for inclusion using objective criteria, are always an appropriate starting point. But identifying all potentially relevant studies is only the first step. In order to make appropriate inferences from this compilation, rather than proceeding to simply lump them together to generate a pooled result, the next step should be to examine the methods to look for important differences in the way the studies were done. Rather than assuming (or pretending) they all used essentially the same methodology to measure exposure, disease or other parameters, which is almost never the case in epidemiology, we need to examine each study to assess all key study features that may influence the findings. The assumption in meta-analysis is that all studies followed the same protocols since that would mean the results differ across studies solely due to random error. However, if they differ in other important ways (*e.g.*, quality of exposure assessment, vulnerability to confounding), then it would not make sense to produce a weighted average of the findings. In fact, studies often differ markedly from one another in how well they measure exposure and disease, which potential confounders are addressed, the underlying susceptibility of the population, and other features. To ignore those significant differences and generate a weighted average is misleading, akin to a weighted average of apples and oranges. The algebra can be done, of course, but that does not make it a meaningful statistical product. As discussed below, we can learn a great deal from the differences in methods and how those differences relate to the study results rather than glossing over the differences, pooling the data and interpreting the product as the most informative measure of the causal effect.

Another feature of meta-analysis that can be problematic concerns the extraction of selected results from each study in an effort to produce comparable findings. Epidemiologic studies typically produce an array of findings that collectively contribute to an understanding of the relationship between exposure and disease. They may use different exposure metrics, control for different factors, or apply different analytic methods, all within the same study. Scrutiny of the array of findings *within* a study is often of great value and that ability is lost when only isolated results are considered. The decision about which findings to extract from each study also leaves room for inconsistent, arbitrary selection across studies, i.e., cherry-picking. The desire for simplicity is understandable and it is simpler to summarize an entire study with one number, but the cost of that simplicity can be a substantial loss of information and a misleading result.

Finally, there is a concern with overinterpreting the product of meta-analyses. Often the final estimate of the association between exposure and disease is based on many studies and appears to be quite precise because of the huge size of the collective set of studies. In presenting this number as a distillation of what the research tells us (losing valuable information along the way), recipients of that information may be led into an overly confident interpretation of how informative this number is. After all, combining many studies to produce a single number seems quite reliable as an indicator of the causal effect, far better than any one study or subset of studies could provide. But such a result tells us nothing about the quality of the studies, only something about how big they are, and we lose information on the quality of the component studies, individually and collectively. All these issues are effectively set aside to focus only on random error, with study size alone determining how heavily to weight individual studies in calculating the pooled estimate. Although it is often quite precise, the pooled estimate can be precisely wrong as a measure of the causal effect. It is a challenge to substitute a more accurate, nuanced interpretation that acknowledges meaningful differences among studies for one that is simple and therefore appealing, but the role of an expert in epidemiology is to delve into the complexity of multiple studies of differing quality and reach an informed judgment. If it could be done well by an algorithm, that would make things much easier, but it cannot be done without a careful assessment of the individual studies.

IV. INFORMATIVE STRATEGIES FOR INTEGRATING EVIDENCE ACROSS STUDIES: CATEGORIZE METHODS

A more informed approach to combining information from multiple studies of the same topic is to begin by taking stock of the full range of methods that have been used in these evaluations. Rather than presuming (or pretending) they are essentially all the same, as is done in meta-analysis, we begin by looking at key features of the research:

- What are the different ways that exposure has been evaluated? Was exposure measured directly or were exposures inferred based on location of residence or occupation?
- How was the health outcome identified? Were medical records reviewed, were the health outcomes self-reported or were they gleaned from death certificates?
- What major potential confounders were controlled? Was smoking history considered in some studies but not others? What about socioeconomic status, obesity, or access to medical care?

Even a preliminary look at the extent of the published literature will provide the menu of approaches that have been applied to address any given topic. In this initial assessment, no value judgments are made, just a cataloguing of how the research was performed. The product of this step is to classify the studies along several different axes that are important determinants of how valid the study's results are likely to be, for example, those that assess exposure with biomarkers (strongest), self-report (weaker), or based on location of residence (weakest).

The next step is to consider the implications of those different methods with respect to the validity of the research and if biases are likely to be present, the likely direction and magnitude of those biases. For these purposes, the evaluation would focus on key study attributes and the relative quality of the different methods. In many cases, the different methods can be rank-ordered as "better" or "worse" but it is helpful even just to note that they are "different" from one another. Having put the different methods into bins, a careful assessment is needed to determine the implications.

On one level the question is simply whether a specific approach is or is not likely to generate an accurate estimate of the causal effect of exposure on disease, i.e., how good or bad is it? But a deeper look is needed to ask how it may be inaccurate, focusing on the implications for the study results. At minimum, we want to learn from assessing these methodological differences if a given study is more likely to overstate or understate the association relative to the true causal effect. Beyond the direction of error that is likely to result from imperfections in the method, we want to know how big of an effect it is likely to have, whether it would be expected to produce only small deviations between the measure of association and the causal effect or could result in a substantial amount of bias. To make these assessments we may have to look beyond the study that was performed.

For example, if we want to know how effectively they have measured the health outcome, we look for validation studies comparing the approach used in a particular study with some gold standard evaluation and determine whether the approach is likely to be accurate and whether it is more likely to produce false positives (assigning disease when it is not really present) or false negatives (missing disease when it really is present). This requires consideration of how the disease manifests itself and whether all cases are likely to be identified. Severe conditions with overt symptoms will likely be identified (*e.g.*, emphysema, myocardial infarction) whereas other conditions may be asymptomatic, exhibit symptoms common to many illnesses, or simply more subtle to identify (*e.g.*, autoimmune disease, thyroid disorders). The key issue is to use what we know about the health problem of concern to assess how accurately the study methods correctly identified its presence.

If we are concerned about adjusting for a particular potential confounding factor, we like to know how strong the confounding could be and in which direction—looking to studies that have examined the potential confounding factor as a risk factor for the disease. This is a particular concern when the potential for confounding is substantial such as assessing a possible environmental cause of lung cancer in the presence of cigarette smoking. If smoking is closely associated with the exposure of concern, then failure to adjust carefully may result in a spurious positive association between the exposure and disease. Again, this is based on knowledge of what other factors are related to the disease of interest and considering whether those other causes may be correlated with the one we are interested in evaluating.

We are then ready to place the array of studies into bins based on the different methods, considering each major determinant of study quality. That is, we are not just providing a global estimate of whether the study was good or bad but rather looking at key characteristics to make assignments based on those considerations, one at a time. We may, for example, look at the different ways exposure was assigned and find some studies using self-report, some using biological markers, and some assigning exposure based on location. For a given confounder, we may have studies that adjusted carefully for confounding and others that did not. The assessment needs to be tailored to the exposure and disease of interest, judging whether the methods used are likely to be effective on a case-by-case basis. Furthermore, the basis for this judgment of study quality needs to be fully explained, that is, why one approach or another to measurement is subject to error or how a given potential confounder would be expected to influence the results.

Combining these categories with an assessment of the quality of the various alternative approaches, we can organize the literature for evaluation. Presumably we will end up with groups of studies (or even individual studies) that have more accurate assessments of exposure or disease, for example, and be able to identify subsets of studies most vulnerable to overstating or understating the association between exposure and disease. Note that all of this can and should be done based on the methods used and what we can predict about how those methods would impact study results based on prior research. This is the exact opposite of cherry-picking based on the results since it provides a reasoned, objective basis for evaluating the research based on the methods that were employed rather than the results that were obtained.

V. INFORMATIVE STRATEGIES FOR INTEGRATING EVIDENCE ACROSS STUDIES: LINK METHODS TO RESULTS

Having organized the studies based on the methods used, grouping them based on the key influences on study quality, the next step is to consider the relationship of the methods to the pattern of results. Review papers often do this to at least some extent and can help to organize and interpret the literature. On a simple level, if the studies can be grouped into those that are better and those that are worse with regard to important features, *e.g.*, those that measure exposure accurately and those that measure exposure poorly, we take note of the pattern of results for those two groups. Obviously, the "good" studies are going to provide a more accurate estimate of the causal effect of exposure on disease than the "bad" studies. This is where we go beyond just noting differences in the study methods but make some judgment about which are likely to be better or worse, ideally based on empirical evidence. If the better studies generate results that are indicative of a positive association, then there is a basis for arguing that a causal effect is more likely to be present. We would not be deterred just because poorer quality studies failed to find such an association. Note that this is quite different than a meta-analysis that lumps them altogether and generates an average across the good and bad studies.

In fact, it may be especially convincing that a causal effect is present if the weaker studies do not find an association since it helps us to understand and explain why this pattern is present. This is far more useful than simply noting the results were mixed or inconsistent. If we understand the basis for the variation in results, then what may superficially appear to be equivocal findings would instead be supportive of an effect of exposure on disease. Note that this would also apply in the other direction—if the superior studies find no association and poorer quality studies that are subject to methodological flaws likely to generate spurious positive findings are supportive of an association, the overall assessment of a causal effect would be that a causal effect has not been found. In this case, mixed results or inconsistent evidence of an association does not mean the studies are inconclusive as is often argued. The results are mixed because the quality of the methods varies across studies and may be much clearer once the studies are grouped based on their quality.

There are several strategic advantages to this approach to summarizing a body of research. First, it is objective, driven by the quality of the methods and not by a preference for one set of results or another. The methods determine the validity of the results and when a subset of studies is based on high quality methods, it makes sense to trust those results more than the results from weaker studies. Second, the rationale can be readily explained and avoids the impression of cherry-picking results based on the studies that generate preferred findings, whether that is for or in opposition to the presence of a causal effect. The logic should be transparent, starting with the basis for believing that one approach is better than another and why that superior approach is less subject to bias and thus a better estimate of the causal effect. Third, this approach provides a rationale for putting some studies aside when their methods are inferior. Without organizing the literature in this way, it is easy to be distracted by less informative studies that are viewed as part of the pool of evidence on an equal footing with the better studies.

One practical challenge is that individual studies may be strong on one attribute but weak on others, *e.g.*, they measure exposure well but are less thorough in controlling confounding. We may not have the good fortune of a subset of studies that are the best in all respects. By going through the exercise of comparing those that are better and worse on each key consideration, one at a time, we may be able to determine which features really do matter and which ones do not. If grouping the studies based on a feature thought to be important turns out not to affect the results, we may decide that it's not so important after all. If we grouped studies into those that controlled potential confounding thoroughly and those that did not do so and found the results were similar for the two groups of studies, we may put that issue aside and focus on more important determinants of study validity.

VI. APPLICATION OF EVIDENCE SYNTHESIS TO LEGAL ISSUES

In communicating the product of evidence synthesis to a non-technical audience, it is important to be able to explain the approach and inferences in a clear way. The essential steps in this assessment are straightforward: (1) Identify the key methodologic features that determine study quality, *e.g.*, control of confounding, accurate assessment of exposure; (2) Group studies based on quality into those that are more and less likely to provide an accurate indication of the causal effect; (3) Examine the results of those studies to determine whether the high quality studies provide a clear or consistent pattern of results that would override the results from the inferior studies.

When there are multiple studies of the same topic and they do not all point in the same direction, which is often the case, the immediate, intuitive response may be to infer that we simply do not know whether there is an effect of exposure on disease. If experts claim that the results are clear because the methodologically stronger studies do show a consistent pattern and the weaker studies only add noise, it may come across as cherry-picking. The rationale for trusting some studies more than others needs to be clear. In fact, when the research findings are mixed, it is important not just to explain why the stronger studies provide more valid information but also to explain why the weaker studies may be misleading, i.e., why they are believed to be subject to bias. This has to be done in a way that is logical to be compelling—starting with the key issues that distinguish good from bad studies and then grouping the studies based on those methods and finally to the results of the studies of varying quality. As we will discuss in Chapters X and XI, in pretrial motions and hearings on causation, some judges attempt to do a deep dive into the studies supporting the competing positions of the parties. Explaining this methodology to the court is essential in preventing a trial judge with insufficient knowledge and experience in this field from becoming confused and misled.

The features that make the good studies good and the bad studies bad needs to be articulated clearly. There is a need to go beyond the message "the poorer studies may be wrong" to explain the nature of the biases to which they are susceptible. For example, study A is not reliable because it measured exposure based solely on self-report, which is known to be inaccurate. Therefore, the study is likely to have failed to detect any adverse effects that are truly present and consequently, does not provide evidence against an association being present. Or, as another example, a key confounder that is positively related to exposure and the disease in question was not controlled so the positive results from a particular study are likely due to confounding by that other exposure. An example of this would be a study looking for an increased incidence of lung cancer due to an environmental chemical exposure that is more common among smokers without controlling for smoking history to isolate any effect of the environmental chemical. Biases generally distort results in predictable directions and the credibility of claims of bias should be supported by the patterns of results. Hypothesizing the presence of a bias with a well-defined pathway, examining the data that would support the operation of the bias, and finding that it is in fact influencing the results provides a compelling basis for dismissing some studies and instead relying on others.

Ultimately, the conclusions from assessing a body of epidemiologic research are often tempered with some degree of uncertainty. Matters under legal contention are often arguable, which is part of the reason why they have become legal disputes. Where the evidence is overwhelmingly in favor of a causal effect, or where the evidence fails to provide any meaningful support for a causal effect, the epidemiologic evidence is not likely to become a key point of contention. But the context makes a real difference in that wide area in between those extremes. With strong ancillary evidence, there may be real concern with whether the epidemiologic research helps to support toxicology studies, for example, or weakens the evidence for a causal effect. Lack of definitive evidence does not mean that no conclusions can be drawn, but only that the conclusions will be tempered by uncertainty. Both the general trend in the epidemiologic evidence and the limitations should be provided. Trying to present ambiguous evidence as compelling is not likely to withstand counterarguments. An informed, balanced assessment is both scientifically optimal and, if the process and outcome is communicated clearly, likely to be viewed as most trustworthy in court.

Chapter 8

INTERPRETING NEGATIVE STUDIES

In this chapter we will explain how to evaluate negative studies, considering the role of statistical significance testing and the significance and common reasons for false negative studies that fail to identify effects that are truly present. These considerations provide a framework for making an assessment of whether negative results reliably indicate the absence of an association.

I. INTRODUCTION

"Negative studies," are those that evaluate whether exposure is associated with disease and do not find support for a positive association. However, interpreting the significance of negative studies can be confusing. The very terminology, in which studies are labeled as "positive" or "not positive" conveys the implication that the studies that are "not positive" have fallen short or failed in some way rather than providing a potentially accurate indication of the absence of a causal effect. Even more directly, the myth that "you cannot prove a negative" suggests that you can only prove a positive or fail to prove a positive. Instead of simply asking how positive the evidence is, a better approach is to ask in a more neutral manner what the research indicates regarding the size of the causal effect of exposure in disease. It may suggest that exposure causes an increased risk of disease of a given magnitude, a decreased risk of disease, or indicate that it has no impact on the risk of disease. If the study employs high quality methods, then it is guaranteed to produce an accurate estimate of the causal effect. For example, there is no reason to challenge the validity of that result because it happens to generate an estimate that suggests no effect at all. We should be ready to accept the findings from appropriately conducted studies, whatever those findings are, and not prejudge some findings as more credible than others. The credibility of the findings is solely dependent on the quality of the methods.

Therefore, digging deeper into the question of whether studies that find no association offer meaningful evidence against a causal effect being present should be done in exactly the same way as it is for studies that find a positive association. We simply ask whether the measure of association that has been generated by the study accurately reflects the causal effect or whether the measure of association has been distorted by study biases. The key difference in this evaluation for studies that generate evidence suggesting a causal effect and those that do not is in the specific types of biases that might account for misleading results. Some types of bias are expected to overstate the causal effect by inflating the measure of association resulting in spurious positive associations, and those are the focus when a positive association is found. In contrast, there are other types of bias that are likely to understate the causal effect and bias the measure of association towards the null value, and those are the focus when an absence of association is found. In either case, we are asking whether the reported measure of association accurately indicates the true causal effect, but in the case of negative studies, we are asking whether the reported absence of effect is accurately indicating no effect or whether biases may have caused the study to fail to identify an adverse or protective effect that is truly present.

To further scrutinize this cliché that "you can't prove a negative," consider a hypothetical scenario in which a series of studies with strong methods have consistently reported relative risks around 1.0, all indicating no association between exposure and disease. In the sense that all research is subject to uncertainty, it may be said that it does not "prove" something but to the extent that research can do so, this would certainly make a compelling case that there is not a causal effect of exposure on disease occurrence. It could just as well be claimed that you cannot prove a positive either, only assemble evidence that indicates an association is present and that it does not seem likely that there are biases that have generated spurious positive findings. The logic is symmetric for positive and negative studies.

II. STATISTICAL SIGNIFICANCE TESTING AND NEGATIVE STUDIES

One of the more common reasons for classifying studies as "positive" or "not positive" is based on statistical significance tests. A study is called "positive" if the measure of the association generates a p-value of <0.05 and considered "negative" (or "not positive") if the p-value is equal to or greater than 0.05. The statistical significance testing framework is formally defined in that manner: either a positive result is obtained or it is not, without considering the implications of "not positive" as evidence for "no effect." As discussed in some detail in an earlier chapter, the way that formal statistical hypothesis testing works is that the default assumption is that the null hypothesis is true (there is no association between exposure and disease) and after the data have been collected, we ask how often if the study were repeated multiple times would random error lead to the association as large as the one that was in fact found. We then decide to reject or fail to reject the null hypothesis, in a sense asking, "are the results positive enough to declare that an association is present?" rather than, "what do the results tell us about the association between exposure and disease?" This may sound like an arcane distinction based on semantics, but it has real importance to the way that studies are interpreted. The most appropriate, informative use of the data is to simply ask first how well-done the study was and conditional on that, see what the study tells us about the association between exposure and disease. By asking "what does it tell us?" rather than asking "how positive is it?," we have the basis for giving well-done studies that show no association the credibility that they deserve, providing meaningful evidence that no association is present. This more agnostic approach de-emphasizes the formalities of statistical significance testing and maximizes what we can learn from the study about whether exposure is a cause of disease. Random error remains a concern, whether the studies do or do not identify an association, and as previously noted, this uncertainty due to random error is best characterized by the use of a confidence interval. Confidence intervals can be generated around any measure of association, including the null value of 1.0 for the relative risk.

Helping to perpetuate this misconception that "negative" implies "uninformative" is the hunger for positive results on the part of researchers, editors of scientific journals, and perhaps the public at large. Assuming epidemiologic studies are asking questions of importance to real-world judgments, which is always the case in addressing legal matters, then the results are important no matter what they show. Negative results from well-conducted studies should not be viewed as "failing to provide a statistically significant positive finding" since they did not fail at all—they provided valuable information regarding the causal relationship of interest. A study is not worth doing if only positive results would be interesting and useful. If the basis for asking the question that the study addresses is valid (based on hypotheses or prior studies), then a credible negative result is valuable.

As discussed in more detail below, the power of a study, its ability to detect an association, is an important consideration in evaluating the relevance of a negative study. If a study is "underpowered," even if an association is found it may not attain statistical significance simply because the study is too small. In a sense, this situation arises as a result of relying on statistical significance testing which fails to distinguish between finding no association and a positive association that is not statistically significant due to imprecision. One of the considerations in judging how informative the study is, whether an association is observed or not, is how large it is and thus how vulnerable it is to random error. And that is determined strictly by how many people are in the study, which is of equal concern whether or not an association may have been found. Again, a point estimate of the association and a confidence interval around that estimate is more informative in describing the study's findings than simply reporting whether the result was statistically significant.

III. COMMON REASONS FOR SPURIOUS NEGATIVE RESULTS

Just as for positive results, when studies generate findings indicating no association, there are two possible explanations: (1) there may truly be no causal effect of exposure on the development of disease and the study has accurately found that to be the case or (2) there really is an effect of exposure on the risk of disease, harmful or protective, but because of limitations in the study methods (biases), it has failed to accurately indicate that exposure is in fact associated with disease. While the potential reasons for misleading negative results are the same as for any other inaccurate finding, some specific forms of bias are the most common contributors to false negative findings. These are the considerations that should be addressed to judge how persuasive the negative results are as an indicator of no causal effect being present. If these potential explanations can be dismissed, or at least determined to have minimal effect, then the negative results can be interpreted as providing meaningful evidence that there is not a causal effect of exposure on risk of disease. Note that invoking these or other considerations to question whether a null result really means no effect does not make a negative study positive, but rather makes a negative study uninformative.

A. Poor Quality Exposure Assessment

Many exposures of interest are challenging to measure to varying degrees, including environmental chemicals, medications, and diet. This is especially true when we are concerned with long-term exposure that can change over time, so we are often reliant on people's memories or other imperfect sources of historical information. A common criticism of studies of diet and chronic diseases such as cancer or heart disease, is the difficulty inherent in accurate self-reporting of lifelong eating habits, often bringing the retort, "I can't even remember what I had for breakfast!" To the extent that exposure is assigned with substantial inaccuracy, approximating random guesses, the predictable result is that no association with exposure will be found. While there are varying degrees of inaccuracy, if we flipped a coin to assign people as "exposed" or "unexposed" it seems obvious that exposure measured in this way will not be related to anything, including disease outcomes of interest. Short of this extreme way of assigning exposure, there are varying degrees of inaccuracy, with some but not all of the people inaccurately assigned, which would blur the association but not necessarily eliminate it altogether.

For this muting of any association to occur due to exposure misclassification, the errors in assigning exposure need to be the same for those who do and those who do not develop disease. Therefore, when an absence of association is found, there is a need to scrutinize the way that exposure was assigned to determine whether it accurately captured the exposure we are interested in. The same null findings can result from an accurate measure of exposure that really does indicate no effect on the risk of disease and a poorly measured exposure that tells us nothing about whether exposure is causing disease. Proper interpretation of a study calls for making this distinction.

Exposure measurement error can occur in a variety of ways. One common way is for studies of diseases that develop over extended periods of time (such as most cancers), but exposure is only known at a point in time, often at the time of diagnosis. Even assuming the assigned exposure is accurate at the time the disease is diagnosed, this may not be an accurate measure of the important exposure, namely the exposure that occurred over the years or even decades preceding the diagnosis. Cancer registry data includes studies that describe the pattern of cancer occurrence in various areas and time periods, with an interest in finding out whether areas that have been subject to an environmental pollutant, for example, have a higher risk of cancer than areas free of such exposure. When studies attempt to link exposure and the frequency of diagnosed disease at the same point in time and then make inferences about the exposure-disease relationship, they are subject to substantial misclassification that often leads to false negative findings. Some of those who were exposed over long periods of time may well have moved away, and some of those who were diagnosed at the location of interest may have recently moved in. Finding no association between current exposure and diagnosis of disease with a lengthy period of development is therefore not very informative in judging whether a causal effect is present.

B. Poor Quality of Disease Measurement

The exact same principles apply to disease measurement as to exposure measurement. To the extent that people in the study are assigned a disease outcome that is inaccurate, and that inaccuracy is the same regardless of whether they were exposed, the measured association will be shifted towards the null value and understate any effect of exposure on disease that is in fact present. At the extreme, instead of diagnosing the disease or querying symptoms or other indications of disease, if we just flipped a coin to assign people as having versus not having the disease, we would not find an association with any exposure, whether that exposure actually causes the disease of interest.

The ability to accurately determine the presence of disease varies. Some conditions, such as myocardial infarction or certain cancers are identified with nearly perfect accuracy. Other conditions such as migraine headaches or developmental delays in children may be subject to some degree of error, especially in their milder forms. We ask how closely the assigned health outcome (the operational measure) corresponds to the gold standard assignment of disease (the truth), whatever that may be. Conditions that inevitably lead to medical care and a full evaluation will be more accurately identified than those that are milder and may not result in seeking care. Along the same lines, health problems that are more severe and

cannot be ignored will be ascertained more accurately than those that are at the margins of being recognized or reported at all. Just as was the case for exposure errors, we cannot distinguish between studies finding no association because there really is no causal impact of exposure on that disease and studies finding no association because they have assessed disease so poorly.

In many cases, the operational measure of disease occurrence captures some but not all of the events of interest. A common exposure is the use of mortality data to identify the occurrence of a disease rather than having incidence data. Because deaths are more easily identified and documented with death certificates, it is often easier to do studies of those who died from a disease rather than those who developed a disease but survived. While determining disease based on deaths is likely to be fairly accurate for diseases that are usually fatal, such as mesothelioma or pancreatic cancer, this type of study would not accurately identify people in the population who contracted cancers or other diseases that are treatable and not commonly fatal, such as thyroid cancer or non-Hodgkin's lymphoma. Studies that rely on mortality data for diseases that are not typically fatal may well fail to identify an association with exposure that is in fact present because of substantial under-ascertainment of those who have developed the health outcome of interest.

C. Negative Confounding

Confounding occurs when another cause of disease is correlated with the exposure of interest such that the measured association of the exposure of interest with disease is distorted. Examples usually refer to positive confounding in which the confounding factor is positively associated with both risk of disease and the exposure of interest. If we are concerned with an effect of coffee consumption on a disease such as bladder cancer, and tobacco use is associated with coffee consumption, we may observe a positive effect of coffee consumption on risk of bladder cancer when in fact the association is really due to cigarette smoking. We need to remove the impact of cigarette smoking to identify the causal effect of coffee consumption on risk of bladder cancer. Often harmful exposures are correlated with one another in a positive way, with those who are at elevated health risk due to one exposure are typically also at elevated health risk due to other correlated exposures. However, there is no reason that the reverse may not also occur in which the confounding factor, which is correlated with the exposure of interest, is protective of the disease.

For example, we might be concerned about whether consumption of certain kinds of fish which contains PCBs due to water pollution is associated with an increased risk of heart disease. However, the same types of fish which accumulate PCBs also have certain nutrients, long-chain fatty acids, that are known to be protective for the development of heart disease. If we determine the association of PCBs from fish with heart disease, we may identify no effect, a null finding, but this could actually be the result of a harmful effect of the PCBs balanced by a beneficial effect of the fatty acids. While the results may accurately indicate that consumption of these types of fish has no net effect, it does not mean that PCBs do not increase the risk of disease. Only if we can study other sources of PCBs or separate the PCBs from fish from other correlated factors like long-chain fatty acids would we be able to discern the harmful effects of PCBs. The same underlying question should be asked: does the null finding indicate no effect of the exposure or might it result from confounding such that there really is an effect of the exposure of interest that is not observed because a correlated exposure has hidden it?

D. Random Error

As noted above, sole reliance on the results of testing for statistical significance can be misleading as a tool for interpreting epidemiologic studies, particularly those that generate null findings. But more generally, random error is capable of distorting results in either direction, overstating or understating the causal effect of exposure on risk of disease. The smaller the study, the more extreme these random fluctuations in findings can be. When a small study generates imprecise results that are essentially null, relative risks at or close to 1.0, it is a valid criticism to note the uncertainty in that estimate and have less confidence that the negative result really signifies the absence of a causal effect. There are two possible reasons that such studies generate negative results—either they have accurately estimated the true, causal effect (null) or there is a causal effect but random error has generated a misleading result. Logically, small studies could just as readily err in the other direction, indicating a harmful or protective effect when neither is present, making them less informative regardless of the results.

In considering this candidate explanation for negative results, there are tools for informing the judgment of whether the study would have been capable of identifying an effect if one had been present. This is referred to as the statistical power of the study. Even small studies may be capable of detecting a huge effect of exposure on disease. If we are interested in smaller, more subtle effects, it becomes increasingly difficult to distinguish between no effect and one that is too small for the study to find. Larger and larger studies are needed to have confidence in interpreting negative findings as meaningful evidence against an association. The same study with negative results may provide strong evidence against a very large effect of exposure on disease risk, but weak evidence against a more modest magnitude of effect of exposure on disease risk and essentially no evidence against a tiny effect of exposure on disease risk. So, for instance, a negative study may rule out the likelihood that a large percentage of people exposed will develop a particular condition, provide some, but often not compelling, evidence that a significant percentage of people exposed will develop the condition, and no meaningful evidence that a small number of exposed people will develop the condition. If even a small percentage of people could possibly develop a routinely fatal condition from exposure, it would be foolish to rely on a negative study of this type for assurance of the absolute safety of exposure to the substance.

IV. CONCLUSIONS REGARDING NEGATIVE STUDIES

Some of the misunderstanding of negative studies comes from conflating the question, "How informative is the evidence?" with the question, "How positive is the evidence?" It should be emphasized that we do not do studies to generate support for a causal effect but rather to provide an accurate estimate of the magnitude of the causal effect. A magnitude of zero is no less informative or important than some other value that indicates a harmful (when investigating a potential toxin) or protective (when investigating a new drug) effect. The question of what the actual causal effect is should be asked agnostically and evaluated critically, whatever the study's findings may be. The informativeness of the study is determined solely by the quality of the methods and not by the results that are obtained. Well-done studies deliver informative results, poorly done studies are subject to error and can be misleading, especially to judges and jurors.

The main features that distinguish the evaluation of negative and positive studies are the sources of bias that most commonly distort the results in one direction or another. When we ask whether the findings of no association provide strong evidence that there is no effect of exposure on disease, we need to ask specifically about what sorts of methodologic limitations may have generated spurious negative results when in fact there really is an effect of exposure on disease. The primary suspects for introducing bias in these cases are distinctive from those that generate spurious positive results, but the logic and approach to the evaluations are the same. By considering the set of spurious reasons for a study to indicate no association when a causal effect is in fact operating, we either find reasons to doubt the accuracy of the studies or we fail to find biases that could generate spurious negative results, therefore increasing confidence that the study has accurately indicated the absence of an association.

The commonly used Bradford-Hill criteria are predicated on having already found a positive association but are sometimes misused to judge whether such an association is present. They were originally developed to help guide the judgment of whether an established association is likely to indicate a causal effect. Therefore, the considerations were not intended to help evaluate whether the absence of an association accurately indicates the absence of a causal effect. Instead, this chapter offers a series of ways in which spurious negative results may be generated. Considering and dismissing each of these biases that commonly cause studies to fail to detect a causal effect is essentially the mirror image to the Bradford-Hill criteria for evaluating the significance of a positive association. The goal is to determine whether the absence of association can be interpreted as the absence of a causal effect. If each of these candidate sources of spurious false negative results can be discounted, we may conclude that the null findings accurately represent an indication of the absence of a causal effect.

This reasoning can be applied to individual studies but is most useful when there is a larger body of relevant research to be evaluated. In many cases, this will include seemingly incompatible results, with some studies indicative of an effect and others finding no effect. Therefore, the evaluation of the evidence in the aggregate requires an informed weighting of the pertinent studies. An evaluation that only focuses only on the studies with positive findings without accounting for high quality negative studies will lead to an invalid conclusion and should be challenged.

Chapter 9

DISTINCTIVE EXPERTISE OF EPIDEMIOLOGISTS AND HOW TO IDENTIFY IT

In this chapter we describe the distinctive features of epidemiology and skills needed to serve as an expert. We address the Daubert criteria, distinguishing true experts from hired guns, and the needed components of a report from an epidemiology expert.

I. THE FIELD OF EPIDEMIOLOGY

Identifying experts in the field of epidemiology can be somewhat challenging given a number of closely related specialty areas. In addition, there is a tendency for those whose work touches on epidemiology to identify themselves as epidemiologists as a secondary specialty combined with their primary area of focus. Among the many realms adjacent to epidemiology are toxicology, biostatistics, environmental exposure assessment, clinical medicine, pharmacology, and health services research. To be fair, epidemiologists often do the same thing, declaring themselves to have expertise in related areas when in fact they are merely conversant with or familiar with the other fields. True expertise in epidemiology (or other disciplines) requires some combination of relevant training and demonstrated contributions in the application of epidemiologic methods to the study of disease determinants. For application in legal settings, it is necessary for epidemiologists to have command over the methods so they can consider and explain what is often a rather nuanced basis for their conclusions about the research on a given topic.

The blurring of boundaries of epidemiology is to some extent inherent in the nature of the discipline. Epidemiologic reasoning is a technically refined version of common-sense, which is both its strength and a source of vulnerability. What could be more obvious than inferring cause and effect relationships from what we observe around us? We do this all the time, inferring the causes of our allergies, how we caught a respiratory illness, how stress has caused gastrointestinal symptoms, how our diet affected our mood. The problem, of course, is that "common sense" without technical underpinnings and careful scrutiny is often misleading. With an anchor in the science of epidemiology, the art of using epidemiology to address causes of disease is compelling, but without that anchor it is simply individual speculation. We each "know" certain things to be true, but these "truths" are often at odds with one another.

In the modern era, an epidemiologist will almost always have training at the masters or doctoral level in epidemiology or as a central part of a more general program in public health or research methods. Those with training in fields such as statistics, clinical medicine, or toxicology may have sufficient training in epidemiology as well, but having that training is not universal for those specialties. The key to mastery of epidemiology is the combination of the subject matter knowledge and research methods. For example, biostatisticians may have the needed expertise in research methods but may or may not be able to examine issues related to exposure or disease ascertainment with sufficient comprehension. Similarly, experts in exposure assessment (environmental modelers, industrial hygienists) may not have deep knowledge of health outcome assessment, and clinical medicine specialists may be unfamiliar with research methods or the specific concerns in exposure assessment even if they know a great deal about the diseases of concern. To some extent, epidemiologists specialize in being generalists, conversant with all the issues that are needed to make inferences about the causes of disease and capable of using insights from related fields even if they lack deep expertise in any of those fields.

II. DISTINCTIVE FEATURES OF EPIDEMIOLOGIC EVIDENCE FOR CAUSAL INFERENCE

While epidemiology is not the only discipline that addresses general causation, it is often the primary one, frequently in conjunction with toxicology and mechanistic research. In some disciplines, there may be a single definitive study that supersedes all that have come before it and stands as the final word on the issue. That is almost never the case with epidemiologic studies, in part because they are conducted with free-living populations in the real world which cannot be controlled, and there is rarely an ideal. With sufficient resources, a randomized clinical trial can be conducted under the ideal conditions because the investigator has control over the study conditions. In observational epidemiology, there is usually a body of research of varying quality, but quality is defined along multiple axes—quality of exposure assessment, quality of disease assessment, vulnerability to confounding, etc. Studies may excel in some respects and fall somewhat short on others. Thus, a review and explanation of epidemiologic evidence for making causal inferences needs to consider and communicate a more complicated story to a lay audience than may be the case for other lines of evidence. Because there are multiple considerations that lead to a final judgment, the reasoning must be clearly articulated or it will appear to be arbitrary.

A corollary of the need to consider multiple attributes of the research is that interpreters must be able to use imperfect knowledge to make inferences. It is not helpful to say we have less than absolute certainty and therefore cannot draw conclusions—we always have less certainty than would be ideal but can draw reasoned conclusions nonetheless. When reviewers go through a series of studies and dismiss them one by one as having some "fatal flaw," they are failing to do justice to imperfect but nonetheless somewhat informative studies. The challenge is to give each study the credibility or weight that it warrants in contributing to the overall assessment of causality, and it is rarely a weight of zero. Expert judgment calls for discerning patterns in the research, often comparing results across studies with varying types and degrees of strengths and limitations. Again, if the interpretation is not done well, it can seem arbitrary and self-serving, retrofitting the evidence to support some preconceived conclusion. But if it is done well, it is clear and should be compelling in connecting the research and methodologic considerations to a carefully documented conclusion.

III. SKILLS NEEDED FROM AN EPIDEMIOLOGIST

The array of skills needed from an epidemiology expert varies to some extent on the question that has been asked, but there are some generic ingredients commonly required. Since the topics under consideration cover an extremely wide range of exposures and health outcomes, as well as responding to the specific legal issues under consideration, there is a need for epidemiologists to be versatile. It would be rare for the epidemiology expert to have all the needed knowledge of the topic prior to pursuing the task that has been assigned to them. They should have the toolkit for acquiring that knowledge, of course, and that includes the ability to grasp the substantive issues regarding exposure, health outcome, and the nature of the causal hypothesis under consideration. While other experts may well be needed to assess exposure in more detail (industrial hygienists, environmental engineers and modelers, etc.) and health outcomes (clinical specialists such as oncologists or psychiatrists), the epidemiologist must be conversant with all of the key aspects of the scenario of interest that are needed to interpret the epidemiologic research. In that regard, epidemiologists who have a reasonably diverse array of experiences would be preferred to those who have only studied a narrow topic over the course of their career.

Quantitative expertise in biostatistics is needed to evaluate, interpret, and explain the evidence. This is a mandatory component of training in epidemiology and depending on the nature of the research, even basic knowledge may be sufficient. But there are bodies of research that are more demanding and may call for a higher level of expertise to properly interpret the findings. Perhaps most important is the epidemiologist's ability to translate the quantitative evidence in a form that is understandable to non-experts. The statistical information is particularly challenging to translate in a clear and persuasive manner.

Particular legal questions may call for epidemiologists who have specific backgrounds in one realm or another. There are many different ways of classifying specialties in epidemiology, some of which are defined by the exposure or potential cause of interest. These would include environmental epidemiology, pharmacoepidemiology, genetic epidemiology, social epidemiology, and nutritional epidemiology. Others are defined based on the realm of application, such as clinical epidemiology which is concerned with research that has direct applications to clinical medicine. And many are defined based on the disease(s) of interest, including cancer epidemiology, psychiatric epidemiology, reproductive epidemiology, cardiovascular disease epidemiology, and neuroepidemiology. While it is not mandatory that the epidemiology expert have familiarity with the particular question being asked in a legal case, there is a real benefit if the expert has some experience with the broader realm in which it falls. If the dispute involves an environmental agent and cancer, there is a distinct advantage in having experience in the general area, for example.

IV. *DAUBERT* FACTORS IN SELECTING AN EPIDEMIOLOGY EXPERT

As will be discussed in Chapter 11, when challenged under *Daubert*, courts do consider whether an expert was familiar with the subject matter of his or her testimony prior to arriving at the opinion being offered, or whether that opinion was arrived at entirely for purposes of litigation. Thus, although it is not required that the precise question at issue was the focus of the expert's earlier research, it is extremely helpful if it was. At the other end of the spectrum is an expert that has never studied anything closely related to the topic of his or her opinion and only first began to address the issue when hired for a particular case. The latter category of expert is at risk of exclusion on a *Daubert* motion and for this reason, everyone with an epidemiology degree will not be viewed equally by the courts.

Because physicians and some other related experts utilize epidemiology in their professional practices, and a medical expert will be required to provide a specific causation opinion, a plaintiff's attorney may wonder whether an epidemiologist is needed at all. A review of Chapter 11, however, will lead to the conclusion that as a plaintiff you will be in a stronger position having an epidemiologist on board than relying on a medical doctor to interpret the epidemiological literature. As the Ninth Circuit observed after the *Daubert* case was remanded from the Supreme Court, "the party presenting the expert must show that the expert's findings are based on sound science, and this will require some objective, independent validation of the expert's methodology."[1] It will generally be easier to meet this burden where the expert interpreting the epidemiological literature is in a profession in which he or she engages in this methodology routinely, rather than occasionally or tangentially. A practicing epidemiologist will also be familiar with all the important concepts that will need to be explained to the court and eventually to the jury, including study design, bias, risk ratios and statistical significance. Medical experts will obviously be familiar with these concepts as well, but articulating them for laypersons may be a challenge to someone not well-versed in the issues and accustomed to explaining them.

[1] *Daubert v. Merrell Dow Pharmaceuticals, Inc.*, 43 F.3d 1311, 1316 (9th Cir. 1995).

Whether representing a plaintiff or a defendant, another factor articulated by the Ninth Circuit in *Daubert* after remand should not be ignored. That an expert testifies based on research he has conducted independent of the litigation provides important, objective proof that the research comports with the dictates of good science. [citation omitted] For one thing, experts whose findings flow from existing research are less likely to have been biased toward a particular conclusion by the promise of remuneration; when an expert prepares reports and findings before being hired as a witness, that record will limit the degree to which he can tailor his testimony to serve a party's interests. Then, too, independent research carries its own indicia of reliability, as it is conducted, so to speak, in the usual course of business and must normally satisfy a variety of standards to attract funding and institutional support. Finally, there is usually a limited number of scientists actively conducting research on the very subject that is germane to a particular case, which provides a natural constraint on parties' ability to shop for experts who will come to the desired conclusion. That the testimony proffered by an expert is based directly on legitimate, preexisting research unrelated to the litigation provides the most persuasive basis for concluding that the opinions he expresses were "derived by the scientific method."[2]

Finding and retaining an expert who has published on the subject area of his or her testimony will provide instant credibility both to the trial judge when deciding a *Daubert* challenge, and to a jury. However obvious this statement may be, accomplishing this is never easy. Some experts are rightfully concerned they will be viewed negatively by their colleagues if they agree to testify in lawsuits involving the subject matter of their research. Moreover, once retained, such experts will be required to disclose their work in legal cases in their future publications, something some are hesitant to do.

Of course, experts who have published on the subject matter at issue can also be effectively cross-examined based on these publications. In the discussion section of every published study, it is common for the authors to objectively present the study's limitations, and, in fact, no epidemiological study is without limitations. Thus, an expert's publications must be scrutinized before a decision is made to prevent unpleasant surprises later on and it is important for the expert to be able to explain in a convincing manner

[2] *Id.* 43 F.3d at 1317.

how research that has some limitations is nonetheless informative and supportive of drawing conclusions.

If an epidemiologist who has published on the subject matter cannot be found, the next best thing is to utilize someone who has conducted and published studies using similar methodology. One of the factors articulated by the courts in determining the admissibility of expert testimony under F.R.E. 702 is whether the expert has arrived at his or her opinion using the same care and diligence as they would a similar undertaking outside of the litigation context.[3] It will be much easier to demonstrate compliance with standard procedures when the expert can testify to a consistent methodology applied in different but similar professional contexts.

One further concept to consider is whether experts from other fields can help support your epidemiology expert's testimony. As discussed in Chapter 4, one of the Bradford Hill criteria in evaluating causation is *plausibility*, i.e., whether there is a clear biological rationale supporting the association demonstrated in epidemiological studies. This can frequently be done using toxicological animal studies or mechanistic research on cell cultures or other biological materials. For instance, a consistent finding in early toxicological studies of PFOA was Leydig cell tumors in mice and rats injected with the chemical. Studies of the C-8 Science Project population identified an increased incidence of testicular cancer, and subsequent studies have provided further support for such an association. Retaining a toxicologist to review the animal studies which investigated the mechanism of development of these Leydig cell tumors can provide important additional evidence of the association in humans you are trying to prove. Conversely, if multiple animal toxicology studies have failed to demonstrate a similarity in effect and identify any mechanism of action, then this lends credence to the epidemiologist who is testifying there is insufficient evidence to support an association.

V. EPIDEMIOLOGY EXPERT VERSUS HIRED GUN

Well-qualified epidemiology experts should have a record of objectivity and maintain their reputation for integrity. But it is not uncommon for individuals to base their professional career as an expert witness in advocacy, arguing in essentially every case that the alleged

[3] *Sheehan v. Daily Racing Form, Inc.*, 104 F.3d 940, 942 (7th Cir. 1997).

cause of disease is in fact culpable (plaintiff's witness) or that none of the suspected causes of disease are true causes (defendant's witness). This predilection is often present independent of the financial gain from serving as an expert witness, since epidemiologists, like everyone else, have an ideology rooted in politics, philosophy, culture, religion, etc. The public may perceive that these individuals are adopting a position on the issue in exchange for payment, but in fact, it is more likely that they were chosen because they already hold the position. In general, those who tend to implicate every agent as harmful to health typically come from the political left and those who never conclude that an agent could cause health harm come from the political right. But regardless of why it occurs, inordinate consistency across topics should raise some concerns. While forceful advocacy may seem helpful, in scientific and professional circles, an expert is likely to be more credible as a neutral interpreter of scientific evidence and leave it to the attorney to employ their assessment in an advocacy role. Just as the attorney is not the technical expert, the epidemiologist is not the advocate or at least should not be.

The appeal of hired guns who will predictably make a forceful case for or against general causation is understandable. They will likely make impassioned, confident, unequivocal expert witnesses in presenting their conclusions. If they are highly experienced, they will have refined their talents in writing reports, being deposed, and testifying in court. In a sense, they are professional expert witnesses, but may or may not be credible as expert epidemiologists. Perhaps most appealing is their predictability—when attorneys seek someone to put forward a preconceived argument, these are the go-to experts. Sometimes attorneys indicate "We need someone to evaluate the evidence on X" but other times they state "We need someone who will say X." For the latter goal, predictability is needed.

There are also arguments against engaging hired guns. Their credibility as independent experts can be challenged based on an inordinately extensive and consistent history of testimony on one side or the other. Judges or juries may put less credence in their arguments because of the suspicion that they are acting as advocates, not experts. The opposing attorneys will likely engage other experts who will challenge the interpretation of the evidence that is put forward if the foundation of their opinion is questionable. A more balanced, informed assessment from the expert may be more helpful to the attorney in preparing for deposition and challenging experts from the other side in court. It is important to understand how the experts on the other side came to their opinions to be able to challenge their conclusions. Even for a lay audience, demonstrating an objective, balanced interpretation of the evidence, including studies that support and those that are counter to the overall conclusion, provides the basis for explaining how the expert has reached their conclusion taking into account the weight of evidence. Done well, that may be more persuasive than presenting the evidence as though there were no ambiguities. Finally, and perhaps somewhat idealistically, experts who have stature and respect within their fields generally deliver more accurate, informed assessments. Their integrity, prominence, and demonstrated neutrality can be highlighted as the basis for valuing their opinion.

Of course, even if an attorney believes that a hired gun expert will make a compelling witness in front of a jury, without credible experience, methodology and supporting studies, such an expert may not ever be permitted to testify. Some epidemiologists who regularly serve as expert witnesses for one side or the other have never conducted or published a single epidemiologic study. Their publications are frequently in non-peer reviewed journals where they make claims for or against causation based upon their critiques of or conclusions drawn from the work of other scientists, sometimes referred to as "re-analysis." Such experts are particularly vulnerable to *Daubert* challenges and even if they survive, can be eviscerated on cross-examination. For these reasons, attorneys screening potential experts in this field should thoroughly vet these experts for past successful *Daubert* challenges and review their publications to be sure they have the experience and credibility necessary to convince a jury that their opinion should be believed.

VI. EXPECTATIONS FROM AN EPIDEMIOLOGY REPORT

There is a great deal of variability in what is asked of epidemiology experts in terms of the scope of the questions and the level of detail needed in the response. Given that the epidemiologist is engaged to address a specific question of legal significance, it is important that the question and range of issues be as clear as possible from the outset. This helps to avoid major errors of omission (points of concern in the case that are not addressed by the expert) and commission (excessive, potentially costly digressions from the questions of interest). While the expert should have the skills to know what information needs to be compiled and evaluated to address the question, clear communication regarding the exact question to be addressed is an essential starting point. Epidemiologists are accustomed to developing evidence for very different purposes, such as scholarly publications including review papers, or as background material for grant applications. These documents written for their peers typically make assumptions about the level of academic training and interests of their audience, using varying amounts of jargon and introducing complexity that would be inaccessible to those outside the field. There are other distinctions, with academic documents often examining questions without the obligation of providing a bottom line other than the stereotypical "more research is needed." In an expert report for legal applications, it is often necessary to make the best judgment possible in light of the available research with no option of suspending judgment—it either is or is not more probable than not that the potentially harmful (or helpful) exposure can cause an increased (or decreased) risk of disease.

Another way in which academic reports may differ from legal documents is the level of certainty required to declare a causal association is present. While epidemiologists typically do not formalize their threshold for the level of certainty required to make such a judgment, it is certainly much higher than 50.1%. Scientific caution and even skepticism are valued attributes to bring to assessments of evidence in technical reports and the expert needs to recognize that a lower level of certainty is required to conclude that an association is more probable than not.

To reach a shared understanding between the attorney engaging the epidemiology expert and what the expert will provide, the questions leading to the engagement are often best expressed in writing even as a brief description of the issue at hand. Following that, a discussion to fine-tune the scope is generally helpful. It may also be helpful for the epidemiology expert to provide an outline for the planned report, which provides another opportunity to refine the request and add or remove sections at an early stage. Assuming the expert starts with a reasonable orientation to the topic they will address, discussing the expected length of the report can help to calibrate the level of detail desired. There is a very wide range in the length of reports, from a one- or two-page synopsis to extremely long, sometimes tedious, detailed reports. Discussing these structural features helps the writer to keep the goal in mind from the outset and prevents the attorney from being surprised weeks or months later when the draft report arrives.

Finally, deadlines for reports or whatever the desired end product may be, should be clear from the outset to avoid crises or rushed products. Another logistical detail to note is the potential need not to share draft materials or email queries which may be subject to discovery. In some cases, the interchange and discussion of draft materials may all need to be verbal, although changes in the Federal Rules of Civil Procedure which have eliminated the need to produce draft reports and other communications have made the collaborative process between attorney and expert less cumbersome.[4]

VII. COMPONENTS OF A REPORT[5]

The contents of an epidemiology expert will vary based on the topic and the particular legal issues in dispute, but there are some generic guidelines that can be offered as a menu to be considered. Not all will be helpful in every situation, and there may well be particular concerns that call for additional components. But to stimulate the thinking that leads to a report outline for discussion by the attorneys and the epidemiology expert, it is helpful to have candidate sections considered for inclusion.

An initial orientation to the field of epidemiology and its methods is often helpful, briefly explaining how research in this field informs judgments. The use of observational studies to address causal inferences is essential to set the stage for applying epidemiologic principles to the issue of concern. The potential sources of error should be enumerated, explaining how each can

[4] *See* F.R.Civ.P. 26(4)(B) & (C).

[5] It is important to advise the expert as to the rules applicable in the jurisdiction where the case is pending. All federal cases require detailed reports, with most courts preventing an expert from testifying to anything not mentioned in the previously served report. *See* F.R.Civ.P. 37(C)(1). State law varies considerably as to what needs to be included in an expert report, and in some states like New York, no expert report is needed, but only a summary disclosure of opinions, grounds for the opinion and factual matter relied upon, all prepared by the attorney. *See* New York CPLR 3101(d)(1).

introduce bias into the studies and how to go about assessing the impact they may have had. The process for establishing causality by eliminating the possibilities that cause spurious associations is important to set the stage for what follows. With this anchor, the ultimate conclusions will be more clearly grounded in science and not seem arbitrary or be difficult to follow.

The special methodologic issues for the topic of interest should also be elucidated, going from the general concepts to their application to research on the topic under consideration. For example, the general discussion of the impact of exposure measurement error in studies of environmental toxicants may be followed by an explanation of why exposure assessment is especially important in addressing the topic at hand, along with a detailed evaluation of how exposure was measured in the relevant research and the assessment of the relative accuracy of various approaches in assigning exposure. In the case of drug exposures, the concern with "confounding by indication" would be introduced, a bias in which it is the condition for which the medication is taken that results in an increased risk of the disease of concern rather than the drug itself. It is important to elucidate the small number of primary considerations for a given topic in detail rather than using a generic checklist such as the Bradford-Hill criteria. The key considerations that make some studies more informative than others and have substantial impact on the overall conclusions vary across topics and need to be explained fully to set the stage for their application.

A clear description of the research that has been done, presented in relatively neutral terms would come next. Depending on the volume and characteristics of the research, it could be a study-by-study summary but more often there are groupings of studies based on the population (*e.g.*, occupational or community exposure), design (*e.g.*, cohort or case-control studies) or key methodologic features (*e.g.*, exposure assessed through measurement or self-report). If there are major and minor contributors, the level of detail should vary accordingly but with an explanation of what makes some of the studies more informative than others. The basic features of the study should be described, which may be done in the form of a table: study population, design, number of participants, key features of exposure and disease ascertainment, potential confounders, main results. The amount of detail would be tailored to the topic and the type of research that has been conducted. The focus here is on the

study methods, which ultimately determine how valid the study results are. In fact, there is no need to know the results to assess study quality and making that clear helps to establish the expert's unbiased assessment of the evidence. Studies should not be valued based on having generated results that support the desired inference but rather because they have been done well and thus provide substantial weight to the final judgment.

The next component is an integration of study methods and results. This is where the methodologic issues bearing on the research are integrated with the specific studies and their results to inform judgment of causality. There is a search for patterns: Do the methodologically stronger studies tend to generate results that are more (or less) indicative of a possible causal effect? Do the studies that are especially susceptible to a particular bias generate results that indicate this bias is present? Since it is common for studies to generate variable results, some supporting an effect and others not, a logical, transparent explanation of how the evidence was interpreted is needed. At the end, the judgment typically requires assigning more weight to some studies than others. Acknowledging there are studies that run counter to the final conclusion but that those are outweighed by other studies helps to demonstrate thoroughness and objectivity. Ideally, this is all communicated in non-technical, accessible terms, revealing the thought process used by the expert who has carefully assessed the research, applied generally accepted methodologic principle, and emerged with a clear interpretation and explanation that a jury or judge can comprehend. It is also helpful to provide simple examples or analogies to help get specific points across. Charts and graphs can be helpful but also confusing to non-scientists. The expert must always keep the audience in mind and avoid using demonstratives that may be commonly utilized at a scientific conference but are inscrutable to the average juror.

The final conclusions directly express the bottom-line judgment of the expert. For this statement, it is important to take the original question posed to the expert into account and address that question explicitly. It is often in the form of "Is it more probable than not…" It may be helpful to briefly recapitulate the key methodologic considerations and perhaps key studies leading to the final assessment. In reaching this punch line, what precedes it should convey a message to the audience. The goal is to have it come across as having been carefully considered, neutrally evaluated, clearly communicated, and ultimately compelling, even interesting. And that requires avoiding it being confusing, arbitrary ("trust me, I'm an expert"), or tedious.

Chapter 10

GENERAL AND SPECIFIC CAUSATION

In this chapter we will review numerous court decisions addressing the issue of causation, explain how courts have differentiated this into two distinct components and comment on some of the pitfalls of trying to reduce epidemiologic evidence to bright line legal tests. The considerations include statistical significance testing, magnitude of association, and the overall support for a causal effect based on the research.

I. INTRODUCTION

The words "causation" and "causality" can mean different things to different people, especially at the intersection between science and law. This is a critically important and challenging issue within the field of epidemiology, and only becomes more complex when applied in the legal setting. In this chapter we will explore how courts have defined and utilized these terms, and specifically, how they have utilized epidemiology and its principles to determine the sufficiency of causation evidence. There are two general contexts in which epidemiology is frequently discussed in court decisions. The first is where the admissibility of a causation expert opinion is challenged either on *Daubert* or *Frye* grounds, or in some states where a separate challenge is made to the opinion's factual foundation. Alternatively, some courts have addressed whether expert testimony found to be admissible on the issue of causation is sufficient to meet the legal burden of proof. The cases discussed below will address epidemiology and causation in both those settings. Chapter 11 will deal more specifically with expert opinion admissibility challenges.

II. EVIDENCE TO ESTABLISH CAUSATION

Courts have generally agreed that proving *causation* in a legal case requires two separate components: *General Causation* and *Specific Causation*. As explained by the New York Court of Appeals in the context of a toxic exposure case (but equally applicable to any putative cause), "[i]t is well-established that an opinion on causation should set forth a plaintiff's exposure to a toxin, that the toxin is capable of causing the particular illness (general causation) and that plaintiff was exposed to sufficient levels of the toxin to cause illness (specific causation)."[1] Some courts, refer to the latter as "individual causation."[2] In the realm of toxic exposure cases, even where general causation has been established—the toxin is capable of causing the disease—specific causation will typically involve an analysis of whether the exposure at issue was sufficient to reach the threshold necessary to cause the particular injury or disease alleged.[3] Such thresholds are also informed by epidemiological studies along with other lines of research (*e.g.*, toxicology). Thus, both general causation and specific causation can, and in most instances do, involve or require epidemiological support. As the Fifth Circuit described, evaluating the sufficiency of causation evidence is: "a two-step process … [f]irst, the district court must conclude there is general causation [and] if it concludes that there is admissible general-causation evidence the district court must determine whether there is admissible specific-causation evidence."[4]

Several Circuits have held that it is not mandatory for general causation to be supported by epidemiological studies, although it is recognized that attempting to establish general causation without epidemiologic evidence makes the task more difficult.[5] In the silicone breast implant litigation, the Tenth Circuit ruled that non-epidemiological evidence alone was insufficient to counter the manufacturer's proffered epidemiological studies finding no proven link between silicone implants and systemic disease. The court explained the importance of epidemiological evidence in this way: "While the presence of epidemiology does not necessarily end the inquiry, where epidemiology is available it cannot be ignored. As the best evidence of general causation, it must be addressed."[6] The lack

[1] *Parker v. Mobil Oil Corp.*, 7 N.Y.3d 434, 448, 857 N.E.2d 1114 (New York Court of Appeals, 2006).

[2] *McClain v. Metabolife Intern., Inc.*, 401 F.3d 1233, 1242 (11th Cir. 2005).

[3] *Parker*, 7 N.Y.3d at 434.

[4] *Knight v. Kirby Inland Marine, Inc.*, 482 F.3d 347, 350 (5th Cir. 2007).

[5] *Rider v. Sandoz Pharmaceuticals Corp.*, 295 F.3d 194 (11th Cir. 2002); *See also*, *Bonne v. ISP Techs., Inc.*, 924, 929 (8th Cir. 2001) ("observing that "there is no requirement 'that a medical expert must always cite published studies on general causation in order to reliably conclude that a particular object caused a particular illness'") (quoting *Heller v. Shaw Indust., Inc.* 167 F.3d 146, 155 (3d Cir. 1999)); *Bendi v. McNeil-P.P.C., Inc.* 66 F.3d 1378, 1384 (4th Cir. 1995).

[6] *Norris v. Baxter Healthcare Corporation*, 397 F.3d 878, 882 (10th Cir. 2005).

of epidemiological support for an expert's causation opinion has been cited as a consideration as to whether the opinion is reliable.[7] In *Ruggiero v. Warner-Lambert*, the Second Circuit upheld the district court's grant of summary judgment on the ground that general causation between the allegedly offending drug, Rezulin, and cirrhosis of the liver had not been established. The district court had cited the absence of any epidemiological support for this connection as one basis for its holding.[8] Similarly, the Eighth Circuit has held that "the absence of epidemiological evidence does not doom [plaintiff's] case" contradicting the conclusion reached by the district court.[9] The court went on to say, however, "[o]f course, epidemiological evidence might have assisted [plaintiff] in establishing causation, and thus, its absence limited the available tools with which she could prove causation."[10]

In prior chapters, we have discussed the various types of epidemiological studies that assess potential causality between exposure to a substance or drug and an adverse outcome. Depending on the type of studies performed, some indicator of relative risk comparing disease among those exposed to those not exposed (*e.g.*, risk ratio or odds ratio) is calculated to quantify the association. Beyond the ultimate findings of increased risk, some courts have considered the presence or absence of a dose-response relationship[11] and also explored the range of the confidence interval and whether that interval extends below 1.0, an indication that the association is not "statistically significant" based on the standard 5% criterion.[12] (*See* Chapter 6).

Another vexing question that has been discussed but not answered consistently, is whether one study showing a statistically significant association is sufficient to support a finding of causation or whether multiple studies are required. Moreover, courts have differentiated between epidemiological opinions based upon conclusions reached by the authors in published peer-reviewed

[7] *McClain v. Metabolife Intern., Inc.*, 401 F.3d 1233, 1252 (11th Cir. 2005).

[8] 424 F.3d 249, 254 (2d Cir. 2005).

[9] *Glastetter v. Novartis*, 252 F.3d 986, 992 (5th Cir. 2001). However, the court did uphold the district court's decision that the expert's causation opinions were inadmissible but for other reasons.

[10] *Id.*

[11] *In re Joint Eastern and Southern Dist. Asbestos Litigation*, 827 F.Supp. 1014, 1042 (SDNY 1995).

[12] *Brock v. Merrell Dow Pharmaceuticals, Inc.* 874 F.2d 307, 312 (5th Cir. 1989).

studies as opposed to opinions expressed by experts based on their own reanalysis of data contained in published studies, i.e., the data was peer reviewed but the new conclusion is formulated by the expert who is testifying. Causation opinions based upon epidemiological data gathered for purposes of litigation are frequently looked at skeptically. While most courts have steered away from getting too deeply into the weeds in interpreting epidemiological studies to determine the sufficiency of the scientific evidence supporting general or specific causation and have merely tried to distinguish between some epidemiological support for an opinion and none at all, others have embarked on detailed analyses of the strengths and weaknesses of the studies proffered in support of a causation opinion, independently determining the scientific value of each study. As indicated in the previous chapters, this is a complex, challenging process even for experts in epidemiology and fraught for those lacking relevant training and experience.

One plaintiff-appellant argued in his brief that the trial judge "traded a judicial robe for a white lab coat in assessing the validity, reliability and 'fit' of the scientific materials relied upon by [plaintiff's] experts."[13]

As Justice Breyer emphasized in his concurring opinion in *General Electric v Joiner*:

> [The requirement that a trial judge act as a gatekeeper under *Daubert*] will sometimes ask judges to make subtle and sophisticated determinations about scientific methodology and its relation to the conclusions an expert witness seeks to offer—particularly when a case arises in an area where the science itself is tentative or uncertain, or where testimony about general risk levels in human beings or animals is offered to prove individual causation. Yet, as amici have pointed out, judges are not scientists and do not have the scientific training that can facilitate the making of such decisions.[14]

[13] *Amorgianos v. National Railroad Passenger Corporation*, 303 F.3d 256, 264 (2d Cir. 2002).

[14] 522 U.S. 136, 140, 118 S.Ct. 512, 520 (1997),

III. REQUIRED MAGNITUDE OF ASSOCIATION

A particularly complex issue encountered in several cases involving the use of epidemiological studies to establish general causation is interpreting the relative risk found in those studies in the context of the burden of proof in civil cases, which requires that causation proof be supported by a *preponderance* of the evidence. Recognizing that an epidemiological study cannot, in and of itself, establish whether any person's particular illness was caused by a particular toxic exposure or drug ingestion (specific causation), courts have grappled with establishing a level of epidemiologically supported increased risk required for an expert's opinion to sufficiently meet the burden of proof on causation. In the Agent Orange litigation, Judge Weinstein referred to the analysis done by Professor Rosenberg comparing the "strong" and "weak" versions of the preponderance rule applied by different courts.[15] The "strong" view requires epidemiological evidence establishing both it was more probable than not (>50% probability) that the substance caused the illness among those who were exposed and medical evidence ruling out other likely causes.[16] For an individual case, to exceed the 50% probability threshold would require a relative risk of 2.0 or greater, comparing those who were exposed to those who were not. The "weak" view, on the other hand, required only an expert opinion backed by statistical support for an association of unspecified magnitude established through the epidemiological literature.

In *Merrell Dow Pharmaceuticals v. Havner*,[17] the Texas Supreme Court engaged in a lengthy analysis of what magnitude of increased risk determined in epidemiological studies of the drug Bendectin, which was alleged to cause birth defects, was required to satisfy the preponderance of the evidence standard. The court used the following hypothetical to explain its reasoning:

> Assume that a condition naturally occurs in six out of 1,000 people even when they are not exposed to a certain drug. If studies of people who *did* take the drug show that nine out of 1,000 contracted the disease, it is still more likely than not

[15] *In re Agent Orange Product Liability Litigation*, 611 F. Supp 1223, 1261 (EDNY 1995) Rosenberg, *The Causal Connection in Mass Exposure Cases: A "Public Law" Vision of the Tort System*, 97 HARV. L. REV. 851, 857 (1984).

[16] *Id.*, 611 F. Supp. At 1262-63.

[17] 953 S.W. 2d 706 (Supreme Court of Texas, 1997)

> that causes other than the drug were responsible for any given occurrence of the disease since it occurs in six out of 1,000 individuals anyway. Six of the nine who developed the condition would be statistically attributable to causes other than the drug (even if the causes are unknown), and therefore, it is not more probable that the drug caused any one incidence of disease. This would only amount to evidence that the drug *could* have caused the disease. However, if more than twelve out of 1,000 who take the drug contract the disease, then it may be *statistically* more likely than not that a given individual's disease was caused by the drug.[18]

Thus, the Texas Supreme Court, and some other courts, have adopted a rule that for epidemiological studies to support an expert opinion on causation, the studies must show a doubling of risk, i.e., an odds ratio, risk ratio or mortality ratio of 2.0 or more.[19]

Other courts have refused to impose such a rigid standard. *In re Joint Eastern & Southern District Asbestos Litigation*,[20] the Second Circuit reviewed the determination of the district court that the plaintiff had failed to present sufficient admissible evidence supporting a finding that asbestos exposure caused the plaintiff's colon cancer. The district court held that causation could be established by a preponderance of the evidence if the epidemiological studies found a risk ratio exceeding 2.0 or through epidemiological evidence with a risk ratio of less than 2.0 but combined with "clinical or experimental evidence which eliminates confounders and strengthens the [causal] connection."[21] In this context, the court used "experimental evidence" to include animal studies that demonstrate a mechanism by which the agent can cause the outcome. The district court embarked on a detailed analysis of all of the epidemiological studies and determined that plaintiff's evidence "establishe[d] only the conclusions that the association between exposure to asbestos and developing colon cancer is, at best, weak, and that the consistency of this purported association across the studies is, at best, poor."[22] The issue before the Second Circuit was not whether the expert causation testimony submitted by plaintiff was admissible, but whether the evidence was sufficient to uphold the jury's finding that causation had been established by a preponderance of the evidence.[23] The district court had determined that the majority of studies relied upon by the plaintiff's expert in linking colon cancer to asbestos exposure had risk ratios of between 1.0 and 1.5, which the court determined were "statistically insignificant."[24] The court found fault with the methodologies employed in other studies that had higher risk ratios and contended that there was insufficient consistency across studies to contribute to the sufficiency of plaintiff's proof on causation.

[18] *Id.*, 953 S.W.2d at 717.

[19] *Id.*; *DeLuca v. Merrell Dow Pharms., Inc.*, 911 F.2d 941, 958 (3d Dir. 1990); *Hall v. Baxter Healthcare Corp.*, 947 F.Supp. 1387, 1403 (D. Or.1996).

[20] 52 F.3d 1124, 1134 (2d Cir.1995)

[21] *In re Joint E. & S. Dist. Asbestos Litig*, 827 F.Supp. 1014, 1030 (S.D.N.Y 1993).

[22] *Id.* at 1041–1042.

The Second Circuit reversed, holding the district court cited "no authority for its bold assertion that [risk ratios] of 1.5 are statistically insignificant and cannot be relied upon by a jury." The court held that it was far preferable to instruct the jury on statistical significance and to let the jury decide whether studies have any significance.[25] The Circuit Court also held that the district court provided no justification for the wholesale rejection of the studies that actually did exceed its adopted minimum risk ratio of 1.5, an example of Judge Breyer's admonition that "judges are not scientists and do not have the scientific training that can facilitate the making of such decisions."

In *Wright v. Williamette Industries*, the Eighth Circuit reversed a judgment entered in favor of a plaintiff offering a more qualitative standard. "It is therefore not enough for a plaintiff to show that a certain chemical agent sometimes causes the kind of harm that he or she is complaining of. At a minimum, we think that there must be evidence from which the factfinder can conclude that the plaintiff was exposed to levels of that agent that are known to cause the kind of harm that the plaintiff claims to have suffered."[26]

[23] *Id.* at 1043.

[24] *Id.* at 1042.

[25] 53 F.3d at 1134; *see also Allen v. United States*, 588 F.Supp. 247, 418–19 (D.Utah 1984) (explicitly rejecting the greater than 50% standard of causation in connection with statistical evidence), rev'd on other grounds, 816 F.2d 1417 (10th Cir.1987); *Grassis v. Johns-Manville Corp.*, 248 N.J.Super. 446, 591 A.2d 671, 674–76 (App.Div.1991) (holding that trial court erred in precluding opinion testimony based on epidemiological studies showing relative risks of less than 2.0).

[26] 91 F.3d 1105, 1107 (8th Cir. 1996).

IV. INTERPRETATION OF STATISTICAL TESTS AND CONFIDENCE INTERVALS

Confidence intervals from epidemiological studies have also been discussed in several court decisions.[27] To review, a confidence interval provides a range within which the true value is expected to fall 95% of the time. The larger the study is, the narrower the confidence interval will be, providing greater certainty regarding where the correct measure of the association is likely to be. The confidence interval gives information on how likely it is that the true value of the relative risk is 1.0, with the observed elevated risk a product of random error. As also discussed in Chapter 6, an odds ratio or risk ratio of 2.1 in one study, may be more supportive of the existence of a causal relationship between an exposure and an illness than higher risk ratio of say 4.1 if the first study is sufficiently large so as to produce a tight confidence interval of say 1.8–2.4, whereas the second study's confidence interval is much larger, *e.g.*, 1.1 to 7.1. Needless to say, the subtleties in the application of confidence intervals in evaluating the sufficiency of causation proof is above the technical expertise of most courts. Yet some have analyzed the studies down to this level when evaluating the epidemiological proof presented in support of an expert's causation opinion.

The more straightforward question posed by confidence intervals is whether a study with a confidence interval that has a lower boundary below 1.0 can be considered at all. This is just another way of expressing a statistical test using the conventional p-value of 0.05: if $p<0.05$, the confidence interval will not contain the null value of 1.0. Some courts have refused to consider studies with confidence intervals that dip below 1.0, while others have refused to exclude them entirely and considered this factor only as a weight issue rather than an admissibility issue.

V. SUFFICIENCY OF EVIDENCE OF CAUSATION

As discussed in prior chapters, epidemiologists rarely if ever rely on only one study to support their conclusions given the inherent fallibility in any single piece of evidence. Typically, multiple studies replicating findings are required and frequently a meta-analysis or pooled analysis of multiple studies is done to aggregate the data to improve confidence in the result. (*See* Chapter 7 for a discussion of meta-analyses). If data from multiple studies are evaluated together, should the data from studies with confidence intervals that fall below 1.0 be entirely excluded from consideration? Epidemiologists would argue many such studies should not be excluded and one should use all informative studies to assess the strength of support for a causal association., But rigid application of the holdings of certain courts would lead to a different conclusion, excluding consideration of such studies and resulting in a loss of meaningful data.

DeLuca v. Merrell Dow[28] a pre-*Daubert* opinion, discusses statistical significance as well as the difference between the conclusions of the authors of epidemiological studies and expert interpretation of the data underlying those conclusions in supporting or refuting causation. In yet another case involving the medication Bendectin and birth defects allegedly caused by a mother taking this drug while pregnant, the district court determined that the opinion of the plaintiff's expert was inadmissible and granted summary judgment for the defendant. The plaintiff's expert had analyzed data from various epidemiological studies and come to different conclusions than the authors of that study (referred to as a reanalysis of the data, or at least a re-interpretation). Moreover, the expert contended that using a 95% confidence level ($p<0.05$) is not magical but is simply a standard convention among epidemiologists. Bolstered by an article by Professor Kenneth Rothman of Boston University School of Public Health, this expert contended that using this standard confidence interval for type one errors (rejecting the null hypothesis when it is in fact true) can result in high risk of a type two error (failing to reject the null hypothesis when it is in fact false). Analyzing the data from these various studies using a higher p-value, Plaintiff's expert reached different conclusions and found support for an association between Bendectin and the birth defects the plaintiff claimed.

The Third Circuit rejected the premise that admissibility under Federal Rule of Evidence 702 required an expert's conclusions to be published and peer-reviewed.[29] It further rejected the argument that any study with $p> 0.05$ should be totally disregarded, stating "The fact that a scientific community may require a particular level of

[27] Confidence intervals were explained at length in Chapter 6.

[28] 911 F.2d 941 (3d Cir. 1990).

[29] *Id.* at 954.

assurance for its own purposes before it will regard a null hypothesis as disproven does not necessarily mean that expert opinion with somewhat less assurance is not sufficiently reliable to be helpful in the context of civil litigation."[30] The court also highlighted the dangers of the district court relying upon decisions from other courts on whether Bendectin could cause birth defects, where it is impossible to evaluate whether the record before those other courts was identical to the record before the district court.[31] In the end, the Third Circuit in *DeLuca* reversed the grant of summary judgment to the defendant, but did not make any final rulings of the admissibility of the opinions of the plaintiff's expert. Rather, the Court remanded the case to the district court with explicit guidance on the analysis that should be undertaken to reach a conclusion on this issue. [32]

Another case addressing the sufficiency of epidemiological evidence supporting causation was *Amorgianos v. National Railroad Passenger Corporation* where a bridge painter alleged he suffered neurological injuries from exposure to xylene contained in the paint being sprayed on a bridge plaintiff was painting.[33] The district court held that to establish causation, plaintiff must offer admissible expert testimony regarding both general causation, i.e., that xylene exposure can cause the type of ailments from which Amorgianos claimed to suffer; and specific causation, i.e., that xylene exposure actually caused his alleged neurological problems. In rejecting the opinion of the plaintiff's treating physician on causation as unreliable, the court held the published articles linking xylene exposure to polyneuropathy upon which the expert relied were insufficient because (1) they did not provide evidence that short-term xylene exposure such as plaintiff had in this case caused polyneuropathy; (2) all of the articles involved individuals exposed to a variety of solvents, not solely xylene and (3) all the articles connecting solvent exposure to peripheral nervous system symptoms found evidence of symmetrical polyneuropathy only, not the asymmetrical symptoms complained of by the plaintiff.[34] The Second Circuit affirmed the district court, holding the trial judge's conclusion that the opinions were not sufficiently reliable to establish causation was well within the court's discretion.[35]

In *Knight v. Kirby Inland Marine*,[36] the Firth Circuit came to a similar conclusion. There the plaintiff attempted to rely upon two case-control studies to establish general causation through an expert epidemiologist. Conceding that "[c]ase-control studies are not per se inadmissible evidence on general causation," the court nonetheless determined that the district court had not abused its discretion in rejecting the expert's opinion. The two studies relied upon involved exposure to multiple organic solvents in addition to the one the expert opined caused the plaintiff's Hodgkin's lymphoma. Neither study came to any definitive conclusions, stating only that it was "possible that exposure to organic solvents may promote the development of Hodgkin's disease…".[37] While recognizing that "in epidemiology hardly any study is ever conclusive and we do not suggest that an expert must back his or her opinion with published studies that unequivocally support his or her conclusions" the court determined that it was within the discretion of the district court to refuse to admit the expert epidemiologist's opinion.

Another issue that has been discussed in several cases is whether general causation proof requires one or more studies showing that substance X increases the incidence of outcome Y, or whether studies showing that substance X can cause a biological reaction similar to outcome Y is sufficient. In *Kennedy v. Collagen Corp*,[38] the district court had rejected the testimony of plaintiff's expert on the ground that there were no epidemiological or animal studies linking the offending substance (Zyderm) to atypical systemic lupus erythematosus (SLE). The Ninth Circuit reversed, stating "The fact that a cause-effect relationship between Zyderm and lupus in particular has not been conclusively established does not render Dr. Spindler's testimony inadmissible." The court pointed out that the district court failed to consider studies relied upon by the expert that demonstrate that Zyderm could induce autoimmune reactions. SLE is an autoimmune disease.

An important distinction must be drawn between the results of clinical trials vs. studies of occupational or environmental exposure

[30] *Id.* at 957.

[31] *Id.* at 953.

[32] *Id.* at 959.

[33] 303 F.3d 256 (2nd Cir. 2002).

[34] *Id.* at 270.

[35] *Id.*

[36] 482 F.3d 347 (5th Cir. 2007).

[37] *Id.* at 353.

[38] 161 F.3d 1226, 1229–1230 (9th Cir. 1998).

to toxins. In the former, it is accurate to say that the control group will not have been exposed to the drug at all. Thus, comparing the incidence of disease in the exposed population (given the drug) to the unexposed control population (given a placebo) is truly an all-or-nothing proposition. In the case of exposure to asbestos and the development of colon cancer, the control group does not have zero exposure to asbestos because everyone has some background exposure. In situations like this, the risk ratio may be stated as a comparison between the more exposed groups and the least exposed group, or background exposure group. This type of risk comparison is meaningful, but does not allow for the simplistic approach applied in *Havner* where the Texas Supreme Court held that a risk less than 2.0 did not meet the more likely than not standard. The relative risk depends on exactly what the exposure levels are in the two groups being compared.

For instance, in a study of the association between PFOA exposure and kidney cancer, previously collected blood samples were used to identify over 300 cases of kidney cancer and an equal number of controls from the same study group. These blood samples were tested for PFOA and grouped into quartiles of exposure.[39] The risk of kidney cancer in the top three quartiles was then compared to the incidence in the lowest quartile to arrive at relative risk. In this example, the relative risk measures the difference in risk against the lowest exposed group, not against a group without any exposure as in a clinical trial. Accordingly, the use of a relative risk threshold of 2.0 to equate to what is more likely than not as was done in *Havner*, could not be justified using the court's logic in *Havner*.

As mentioned at the outset, proving causation requires proof of two components, general and specific causation. Many courts have approved use of a differential diagnosis methodology to establish specific causation.[40] Using this methodology, an expert "rules in" all plausible explanations that could cause an illness and then "rules out" the least plausible causes until the most likely cause remains.[41]

[39] Shearer, et al., *Serum Concentrations of Per- and Polyfluoroalkyl Substances and Risk of Renal Cell Carcinoma*, JNCI: Journal of the National Cancer Institute, Volume 113, Issue 5, May 2021, pages 580-587, https://doi.org/10.1093/jnci/djaa143.

[40] *Turner v. Iowa Fire Equip. Co.*, 229 F. 3d 1202, 1208 (8th Cir. 2000).

[41] *See gen.*, *Westberry v. Gislaved Gummi, AB*, 178 F. 3d 257, 266-267 (4th Cir. 1999).

However, for a specific causation opinion based on a differential diagnosis methodology to be admissible, general causation must be established for whatever the most likely cause turns out to be. In *Glaustetter*, the court concluded that there was insufficient scientific support for an opinion that the drug Parlodel, a drug given to stop postpartum lactation in woman who didn't want to breast feed, could cause a hemorrhagic stroke. Therefore, the experts' conclusion that it was the most likely cause of plaintiff's stroke using a differential diagnosis methodology was found to be inadmissible.[42]

VI. CONCLUSIONS

Generalized conclusions are difficult to discern from a review of the cases that have addressed general causation and epidemiology. But a few can be identified. First and foremost, it is extremely helpful for a plaintiff to have epidemiological evidence supporting his position on general causation and detrimental if such evidence is lacking. It is easier for a plaintiff to establish general causation where her expert relies upon multiple studies with risk ratios over 2.0 and confidence intervals that do not dip below 1.0. Conversely, defendants have been successful in defeating general causation in situations where there are few studies and those that exist do not meet these criteria. However, there are no bright lines and each case must be evaluated on its own merit.

[42] *Glastetter v. Novartis Pharmaceuticals Corp.*, 252 F.3d 986, 992 (8th Cir. 2001).

Chapter 11

DAUBERT AND *FRYE* CHALLENGES AND OTHER PRE-TRIAL MOTIONS INVOLVING EPIDEMIOLOGY

In this chapter we will review numerous court decisions interpreting epidemiologic evidence in pre-trial motions, particularly motions to exclude expert testimony under the Daubert and Frye doctrines, and discuss how epidemiology relates to class certification.

I. INTRODUCTION

In Chapter 10, we discussed the use of epidemiologic evidence to draw conclusions regarding causation, both general and specific. Because many of the cases discussed were opinions on *Daubert* challenges, there will be some overlap between that chapter and this one. The focus here will be more on the advent of the *Daubert* challenge and the factors that are applied by courts to determine admissibility of opinions based on epidemiology. This chapter will also consider the other doctrine under which expert methodology can be challenged in some states pursuant to *Frye v. United States* and other variations on the same theme where epidemiology is involved in determining whether an expert's opinion has sufficient foundation to be admissible in states where *Daubert* has not been adopted. We will also touch briefly on the use of epidemiology in some recent motions for class action certification seeking medical monitoring damages.

II. *DAUBERT* AND *FRYE*

In Chapter 10 we discussed the long-contested litigation over the relationship between birth defects and the drug Bendectin, which was prescribed from the late 1950s through 1982 to pregnant women to reduce symptoms of morning sickness. The *Daubert* era in federal litigation owes its genesis to Bendectin. Jason Daubert and Eric Schuller and their families sued Merrell Dow, Bendectin's manufacturer, in California state court alleging that the serious birth defects they had suffered, specifically limb reduction defects, were caused by their mothers' ingestion of this medication during pregnancy. The case was removed from California state court to federal court on the jurisdictional basis of diversity of citizenship. Merrell Dow moved to exclude the opinions of plaintiffs' experts under the then existing rule in the federal courts and many state courts as well, the doctrine established in *Frye v. United States*.[1]

The *Frye* test had its origin in a terse decision from the D.C. Circuit Court of Appeals in 1921 concerning the admissibility of evidence derived from a systolic blood pressure deception test, a crude precursor to the polygraph machine. What would become known as the *Frye* test required that to be admissible, scientific evidence had to be *generally accepted* within its relevant scientific field or technical discipline. In establishing this rule that held sway in both state and federal courts for over 70 years, the court explained:

> Just when a scientific principle or discovery crosses the line between the experimental and demonstrable stages is difficult to define. Somewhere in this twilight zone the evidential force of the principle must be recognized, and while courts will go a long way in admitting expert testimony deduced from a well-recognized scientific principle or discovery, the thing from which the deduction is made must be sufficiently established to have gained general acceptance in the particular field in which it belongs.

Plaintiffs' three experts in *Daubert* had relied upon three different types of scientific proof to support their opinions: reanalysis of data from prior epidemiological studies; animal toxicology studies; and the chemical similarity of Bendectin to other teratogens (chemicals already known to cause birth defects). The district court applied the *Frye* test and held that the experts' opinions linking Bendectin to the infant plaintiffs' birth defects were inadmissible because a causal link between Bendectin and limb reduction defects was not generally accepted in the scientific community.[2] The Ninth Circuit upheld this ruling and the grant of summary judgment thereafter due to lack of proof of causation.[3] It was in this setting that the Supreme Court adopted the new criteria for evaluation of the admissibility of expert testimony, which has thereafter been generally referred to as the *Daubert* standard.[4] In adopting this new standard the Court charged trial judges with the responsibility of acting as gatekeepers allowing only reliable expert opinions to reach the jury.

[1] *Frye v. United States*, 293 F. 1013, 1014 (D.C. Cir., 1923).

[2] *Daubert v. Merrell Dow Pharmaceuticals*, 727 F. Supp. 570 (S.D. California, 1989).

[3] 951 F.2d 1128, (9th Cir. 1991).

[4] *Daubert v. Merrell Dow Pharmaceuticals*, 509 U.S. 579 (1993).

The basis for the Supreme Court's decision to abandon the seventy-year-old *Frye* doctrine was the legislative adoption of the Federal Rules of Evidence in 1975, and specifically F.R.E. 702. Ironically, the Court's *Daubert* decision was intended to liberalize the approach to the admissibility of scientific evidence from what it considered a rigid *Frye* standard. While *Frye* may have been rigid, it was also relatively easy to apply. Applying *Daubert*, however, has proved to be anything but, and has led to significant additional litigation within cases and inconsistent outcomes as judges' grapple with evaluating complex scientific evidence well beyond their knowledge base.

Daubert set forth a non-exclusive checklist for trial courts to use in assessing the reliability of scientific expert testimony. The specific factors articulated by the *Supreme* Court are: (1) whether the expert's technique or theory can be or has been tested—that is, whether the expert's theory can be challenged in some objective sense, or whether it is instead simply a subjective, conclusory approach that cannot reasonably be assessed for reliability; (2) whether the technique or theory has been subject to peer review and publication; (3) the known or potential rate of error of the technique or theory when applied; (4) the existence and maintenance of standards and controls; and (5) whether the technique or theory has been generally accepted in the scientific community.[5]

Because the order and judgment before the Supreme Court in *Daubert* was based on the *Frye* standard, the case was remanded for further proceedings. On remand, the Ninth Circuit determined it could apply the new standard without any further hearings in the district court. As the first Circuit Court to attempt to interpret and apply the new *Daubert* standard, Judge Kozinski provided this preface to the task at hand:

> Our responsibility, then, unless we badly misread the Supreme Court's opinion, is to resolve disputes among respected, well-credentialed scientists about matters squarely within their expertise, in areas where there is no scientific consensus as to what is and what is not "good science," and occasionally to reject such expert testimony

[5] *See* F.R.E. 702 Advisory Committee Notes.

> because it was not "derived by the scientific method." Mindful of our position in the hierarchy of the federal judiciary, we take a deep breath and proceed with this heady task.[6]

The Ninth Circuit then interpreted the Supreme Court's *Daubert* decision to require the trial court to first determine whether the challenged opinion has a valid scientific basis. As the Court put it, "... to analyze not what the experts say, but what basis they have for saying it."[7] As mentioned above, one of the three groups of experts the *Daubert* plaintiffs relied upon to prove that Bendectin caused their birth defects were epidemiologists who reanalyzed published data from studies done on Bendectin, coming to different conclusions from those of the studies' authors who found no evidence of a causal association. In judging the scientific validity of their methodology in doing so, the court pointed to two important factors. The first was "whether the experts are proposing to testify about matters growing naturally and directly out of research they have conducted independent of the litigation, or whether they have developed their opinions expressly for purposes of testifying." The court concluded that the former had various checks on its validity within the scientific community while the latter invited bias. Similarly, the second factor, the peer review process for a scientific publication was deemed by the court to provide some scientific validation of the methodology applied. Presumably, a study is unlikely to be published in a reputable peer-review journal if its methodology is badly flawed.

The plaintiff's epidemiology experts in *Daubert* had neither of these factors supporting their opinions. They had not done independent epidemiological research on Bendectin. Although the data they reanalyzed was indeed published, their reinterpretation of that data was never subjected to peer review. Even admitting these flaws, the Ninth Circuit signaled it would have granted plaintiffs' request for a remand but for the court's interpretation of the second prong of the *Daubert* standard under Rule 702, what is often referred to as the "fit" requirement. As the Supreme Court explained: "Rule 702's 'helpfulness' standard requires a valid scientific connection to the pertinent inquiry as a precondition to admissibility."[8]

[6] *Daubert v. Merrell Dow Pharmaceuticals, Inc.*, 43 F.3d 1311, 1316 (9th Cir. 1995).

[7] *Id.*

[8] *Daubert*, 509 U.S. at 591.

The Ninth circuit concluded that because the reanalysis of the Bendectin study data by plaintiff's expert did not produce a risk ratio above 2.0, the opinion did not "fit" the inquiry, because, the court concluded, only a risk ratio above 2.0 provided assurance that the causal relationship was more probable than not and could satisfy the preponderance of evidence standard.[9]

The list of factors provided in *Daubert* by which a district court was to evaluate the reliability of expert opinion evidence was not intended to be exhaustive, and since *Daubert*, additional factors have been articulated by various lower courts including:

- Whether the expert has unjustifiably extrapolated from an accepted premise to an unfounded conclusion—"too great an analytical gap between the data and the opinion proffered;"[10]
- Whether the expert has adequately accounted for obvious alternative explanations;[11]
- Whether the expert has approached the problem as he or she would have approached it outside of the litigation context;[12]
- Whether the field of expertise is recognized as producing reliable results.[13]

An example of a successful *Daubert* challenge to testimony based allegedly on epidemiological evidence is the Second Circuits opinion in *Amorgianos v. National R.R. Passenger Corp.*[14] The plaintiff alleged he suffered neurological injuries as a result of inhalation and dermal exposure to toxic chemicals, specifically including xylene, contained in paints, thinners and primers used where he was working painting a bridge. The district court had ruled that plaintiff's experts could not testify "'on the issue of general causation with respect to Mr. Amorgianos's alleged chronic neurological conditions,' because their opinions were unreliable."[15] Plaintiff called his treating physician, Dr. Moline, at trial, who cited several studies in support of her conclusion that his neurological symptoms were caused by organic

[9] 43 F.3d 1311, 1320–1321. *See also*, discussion in Chapter 10 of this risk ratio threshold analysis.

[10] *General Electric Co. v. Joiner*, 522 U.S.136, 146 (1997).

[11] *Ambrosini v. Labarraque*, 101 F.3d 129 (D.C. Cir. 1996).

[12] *Sheehan v. Daily Racing Form, Inc.*, 104 F.3d 940, 942 (7th Cir. 1997).

[13] *Kumbo Tire Co. v. Carmichael*, 526 U.S. 137, 119 S.Ct. 1167 (1999).

[14] 303 F.3d 356 (2d Cir. 2020).

[15] *Id.* at 260.

solvent exposure at his job site. In upholding the district court's determination that Dr. Moline could not testify as to general causation because "there was too great an 'analytical gap' between the conclusions reached by the authors and the conclusions she draws from their work," the Second Circuit explained:

> With respect to the court's exclusion of plaintiffs' experts' general causation testimony, Judge Trager conducted an extremely thorough review of the scientific literature on which plaintiffs' experts relied, *see Amorgianos*, 137 F.Supp.2d at 192–216. Although this degree of review, while commendable, may not always be necessary to evaluate whether proffered expert testimony is admissible, Judge Trager's evaluation of the fit between the experts' opinions and the scientific literature on which they relied was certainly within the broad discretion afforded to the district court under *Daubert*… Although Dr. Moline relied on a number of published articles in concluding that Amorgianos's xylene exposure caused him to suffer from polyneuropathy, the district court's close analysis of those studies revealed that (1) none of them provides evidence of the neurological effects of short-term xylene exposure; (2) all of the articles involved individuals who were exposed to a variety of solvents, many of which were not contained in the paint Amorgianos used; and (3) all of the articles connecting solvent exposure to peripheral nervous system symptoms found evidence of symmetrical polyneuropathy only, not of the asymmetrical symptoms of which Amorgianos complained. *Id.*

Amorgianos is instructive in that it permits, but does not mandate, a thorough review of the cited scientific literature by the district court to reach a determination on whether the referenced studies actually support the expert's conclusions. Here, the court found, they did not, and it is difficult to take issue with the stated reasons why.

As discussed in Chapter 10, *Daubert* has prompted some courts to dig deeply into the epidemiological evidence proffered in support of a causation opinion, with some judges offering their individual critiques of each study.[16] Because reported decisions are likely skewed in the direction of these microanalyses, it is difficult to generalize.

However, the majority of the case law seems to be derived from the most controversial of the causation battles, where there are either no epidemiological studies supporting a causal association or the studies that have been done uniformly fail to find one. If there are multiple epidemiological studies finding an association backing up an epidemiologist's opinion supporting general causation, there are few successful *Daubert* challenges to be found.

III. STATES WHERE *FRYE* REMAINS THE RULE

Daubert is the controlling authority in all federal courts. In the years since the decision was issued, 41 states have followed along and adopted either the *Daubert* factors themselves or some semblance of those factors in determining the admissibility of expert testimony.[17] Of the remaining states, New York, Washington, Pennsylvania, Minnesota and Illinois continue to utilize the *Frye* standard, while the remaining four, Nevada, North Dakota, Virginia and South Carolina follow neither *Daubert* nor *Frye* and have adopted their own unique rules.

In states where *Frye* remains the rule, some courts have limited their inquiry into the admissibility of expert causation opinions to whether the methodology is "novel." For instance, with regard to whether a chemical was capable of producing a particular medical condition, one court held "plaintiffs' experts relied upon epidemiological studies, which are by no means a novel methodology for demonstrating a causal relationship between a chemical compound and a set of symptoms or a disease."[18]

But although New York continues to follow *Frye*, its approach to the admissibility of expert evidence on causation in particular is not all that different than that applied in the federal courts. The case in New York that began the migration toward a *Daubert*-like approach was *Parker v. Mobil Oil Co.*[19] Although the court declined to adopt *Daubert*, it articulated an approach to addressing scientific causation evidence that has many similarities to that articulated in *Daubert*. Plaintiff in Parker was a service station attendant who developed leukemia. The experts supporting his claim that his illness resulted from his exposure

[16] *See e.g., In re Joint E. & S. Dist. Asbestos Litig*, 827 F.Supp. 1014, 1030 (S.D.N.Y 1993).

[17] *The States of Daubert after Florida*, EXPERT NEWS, Posted July 9, 2019, updated May 6, 2020; https://www.lexvisio.com/article/2019/07/09/the-states-of-daubert-after-florida.

[18] *Jackson v. Nutmeg Technologies, Inc.*, 43 A.D.3d 599, 601 (3d Dept. 2007).

[19] *Parker v. Mobile Oil Corp.*, 7 N.Y.3d 434 (2006).

to gasoline fumes relied upon epidemiological studies showing a causal association between exposure to benzene and the type of leukemia he developed. Because benzene is a component of gasoline, they reasoned that his exposure to benzene through breathing gasoline fumes caused his illness. New York's highest court, the Court of Appeals, rejected the opinions on foundational grounds and affirmed summary judgment for defendant.

The court accepted that there was a scientifically valid causal link between exposure to benzene and leukemia supported by various epidemiological studies, but found that the experts failed to demonstrate that the quantity of benzene to which plaintiff was exposed from inhaling gasoline fumes as a service station attendant was sufficient to establish causation. This analysis is strikingly similar to the "fit" prong of the *Daubert* analysis under F.R.E. 702. The court pointed out various ways the benzene studies could have provided adequate foundation, for example, modeling of the benzene exposure plaintiff was subjected to through inhaling gasoline fumes, but because this was not done, and the experts simply assumed the benzene exposure from gasoline was sufficient to make the epidemiological benzene studies relevant, it deemed the testimony inadmissible for lack of sufficient foundation. Since *Parker*, other decisions from the Court of Appeals and lower courts in New York have excluded expert testimony on similar foundational grounds without relying upon *Frye*.[20]

IV. EPIDEMIOLOGY AND CLASS CERTIFICATION

Historically, epidemiology was rarely considered in class actions. This is because general and specific causation of illness or injury are rarely, if ever, relevant issues in class actions where all class members are required to suffer the same basic injuries caused by the same allegedly tortious conduct. Recently, however, a number of class actions have been brought seeking damages for what is called medical monitoring, a program of medical surveillance for the benefit of people exposed to a toxicant that increases their risk of future illness. Several of these cases have involved drinking water contamination with perfluorooctanoic acid (PFOA) a chemical used in making fluoropolymers that when released into the environment dissolves into groundwater and remains there forever. When PFOA

is ingested or inhaled it remains in the blood for years and can be accurately measured. This has made the study of PFOA associated illnesses much easier than chemicals like trichloroethylene (TCE) which is metabolized in the body and excreted relatively rapidly making exposure calculations problematic. Class certification for medical monitoring classes has been granted by several courts for PFOA exposure based upon epidemiological studies showing adverse health effects from this persistent environmental contaminant.[21] Further discussion of these studies and how they came about is provided in a later chapter.

[20] *Cornell v. 360 West 51st Street Realty*, LLC, 22 N.Y.3d 762 (2014).

[21] *Sullivan v. Saint-Gobain Performance Plastics Corporation*, 2019 WL 8272995 (D.Vt. 2019); *Baker v. Saint-Gobain Performance Plastics Corporation*, 2022 WL 9515003 (W.D.N.Y. 2022).

Chapter 12

DEPOSITIONS, DIRECT AND CROSS-EXAMINATION OF EPIDEMIOLOGIC EXPERTS

In this chapter we will provide practical suggestions for presenting and challenging testimony of epidemiologists in court, both direct and cross-examination.

I. INTRODUCTION

Just as the complexity and nuances of the science of epidemiology present challenges to trial judges in evaluating the admissibility of causation opinions, so, too, does crafting an effective direct examination of an epidemiology expert in support of your case or cross-examining the opposing epidemiologist. This can be a daunting challenge for even the most experienced trial lawyers. In this chapter we will present some suggestions for effectively accomplishing both, again viewed through the perspective of both a trial lawyer who has examined epidemiologists in court and an epidemiologist who has served as an expert witness. Working with an epidemiology expert who is in command of the relevant research and methodologic issues that bear upon its interpretation provides the foundation for the trial lawyer to be most effective in court.

In Chapter 9 we discussed how to recognize epidemiology expertise when retaining an expert to testify in a case drawing on evidence from this field. In the sections below, we will also discuss strategies to maximize the value of such expertise in order to benefit from the engagement of an appropriate expert and exploit the opposition's failure to retain an expert who possesses the background and insights expected from epidemiologist.

II. DEPOSITION STRATEGIES

In all federal cases, and in state court cases in most jurisdictions, parties are permitted to depose opposing experts, and few trial lawyers pass up this opportunity. Like any deposition, the goal when deposing an expert is to obtain admissions you can utilize at trial. But when general causation is the subject matter, you know you will also either be making or defending a *Daubert* motion or both. Thus, your outline of topics to cover when deposing the opposing epidemiology expert should include all of the *Daubert* factors and when preparing your own expert, these factors also must be considered to be able to withstand challenges.

III. EXPERIENCE WITH THE SUBJECT MATTER

As mentioned above, courts and eventually juries will take note of whether an expert has been professionally acquainted with the subject matter of their testimony prior to being hired as an expert. Reviewing the expert's C.V. will typically provide information needed to make this assessment. For example, one epidemiologist deposed in a case had numerous publications listed he had authored. However, none involved original research. When he was not testifying in court, his entire career had been spent publishing non-peer reviewed articles in minor journals criticizing studies that had findings contrary to the economic interests of companies and industry groups who had provided him with financial support. At his deposition he admitted to never having conducted any independent epidemiological research of any kind. This is the type of admission that can provide a basis to challenge the admissibility of an expert's opinion. Even if that fails, it can be used to later challenge his impartiality and credibility in front of a jury.

IV. EPIDEMIOLOGICAL EVIDENCE

The first two factors articulated by the Supreme Court in *Daubert* are whether the expert's theory has been tested and whether it has been subjected to peer review. Thus, a thorough knowledge and understanding of the epidemiological literature on the subject at hand is essential to allow the lawyer to be conversant on both of these topics. Often, assistance from your own expert is helpful in understanding some of the technical aspects of these studies. Prior chapters also provide the necessary background to understand the relevance or lack thereof of negative studies, statistical significance and meta-analyses. Understanding these concepts, and the help of your own expert, will help you devise a plan to demonstrate why your expert's opinion is credible and scientifically based and your opponent's expert's is not.

If there are only one or two studies that support the opposition's position, digging deeply with your expert into those studies to find weaknesses is crucial, especially if you plan on mounting a *Daubert* or *Frye* challenge. In Chapter 3 we discussed various types of bias that can affect the results of an epidemiological study. A review of the concepts identified in that chapter is a good starting point in deciding how to seek admissions from the opposing expert at a deposition. Every epidemiological study has weaknesses, but some are more concerning than others. Well-designed clinical trial designs are the most difficult to attack in a number of respects since exposure is randomly assigned, and most use double-blind methodology so that even the researchers don't know which subjects got the drug being tested and which got a placebo. But workplace and community studies seeking to determine whether a particular agent is capable of causing a particular disease are much more difficult to design and conduct, and thus more susceptible to being challenged.

A frequent area for challenge is how the study handled confounding. For example, cigarette smoking is associated with multiple adverse health outcomes. If a study looking at chemical X finds an excess of lung cancer in the exposed population, unless smoking is accounted for it, the study's value in identifying a causal effect of chemical X on lung cancer is questionable. This is a simple example to understand due to the general understanding that smoking causes lung cancer, but other confounders not considered may be important but more subtle. Again, your expert should be helpful in identifying these and coordinating your direct examination with the points you want to emphasize on cross will provide the jury two opportunities to absorb your arguments.

Disease measurement errors can also be important in studying conditions that are more difficult to diagnose. Early on in the development of epidemiological evidence of the adverse outcomes from asbestos exposure were hampered by the misclassification of mesothelioma as lung cancer. Mesothelioma is now recognized as the signature cancer caused by asbestos which affects the lining of the lung, the pleura. This cancer was rare to non-existent until asbestos became widely used and can take as long as 40 years after the onset of exposure to develop. In the 1950s, numerous cases of mesothelioma were mischaracterized as lung cancer so the causal link between asbestos and mesothelioma took longer to be recognized. If the disease at issue is not one that is either frequently diagnosed or if its diagnosis relies upon finding a set of signs or symptoms rather than some definitive imaging or pathology, the opportunity for mischaracterization increases and the reliability of studies finding or not finding an association may also be called into question.

In addition to evaluating the overall quality of the study methods, it is useful to know whether the limitations of a study are more likely to yield spurious positive results indicating an association or fail to identify effects even if they are present. Previous chapters have considered the anticipated direction of distortion resulting from various methodologic shortcomings.

For example, worker mortality studies often utilize death certificates to determine a cause of death and usually compare a group of workers with presumed exposure to the agent being studied to a presumably unexposed group of workers who are otherwise similar. If the disease at issue is one that is usually untreatable and leads to death rather rapidly, this type of study may be an effective way to determine whether such an association exists. Conversely, if the disease at issue does not lead to mortality in most instances, or if the time between diagnosis and death when it does occur is extremely long, a mortality study is more vulnerable to misleading results, specific failure to identify an adverse effect when one is truly present.

Exploring the details of an epidemiological study with an expert at a deposition can provide helpful information for a challenge to the admissibility of the expert's opinion. These admissions are less likely to be effective with a jury, since these details are complex and difficult for most jurors to understand. However, at times, with clear communications between the epidemiology expert advising the attorney and thoughtful approaches to explaining the evidence, juries can be educated to appreciate the weaknesses of the studies the opposing expert is relying upon.

Another challenge is the "meta-analysis" expert. Some such experts have not designed or authored any peer-reviewed epidemiological studies themselves. Instead, they have pieced together data from a number of studies conducted by others and purport to provide a balanced assessment of causal effect based upon the totality of this pooled data. They often publish these meta-analyses in secondary journals that are not peer-reviewed, adding them to their CV and misrepresenting them as original empirical research.

As discussed at length in Chapter 7, meta-analysis can be a useful tool, but it is often used inappropriately and generates an incomplete and misleading summary of the evidence. This is particularly a problem in epidemiology where the methods are often notably different across studies and simply should not be pooled. If the opposing expert is relying primarily on the results of meta-analyses, it is important to challenge him or her on *Daubert* grounds and use your expert to explain why the meta-analysis tool is inappropriate for the question presented.

V. FIT

As discussed above, even where the epidemiology relied upon by the expert is methodologically sound, the nexus between the studies relied upon and the final opinion is still subject to challenge. Opposing experts can consider the same body of research and come to different conclusions solely because of their interpretation of how that evidence applies to the question at hand. Examples of this previously provided discussed the *Parker* case from New York as well as the *Amorgianos* case from the Second Circuit. The expert in *Parker* relied on epidemiological studies of benzene, but failed to link these studies to gasoline fume exposure sufficiently for the testimony to be deemed admissible. Similarly, the expert in *Amorgianos* relied on general studies conducted on workers exposed to a plethora of chemicals and was unable to identify studies reflecting a causal link between the chemical at issue at the level of exposure at issue in that case.

VI. DIRECT EXAMINATION

Preparing a direct examination of an epidemiologist can be a daunting task. In federal court the strict rules require the expert to testify consistently and within the parameters of his or her report. But while an expert's report may be difficult for a jury to comprehend due to scientific jargon or complex lines of inference, when testifying the expert must be able to explain difficult scientific concepts in plain English, never an easy task. For most scientists, and this certainly applies to epidemiologists, the concept of absolute certainty is a foreign one. The scientific method rewards skepticism and seeks to constantly challenge conventional wisdom. However, in the courtroom, this scientific tendency to hedge one's bets and avoid firm conclusions needs to be managed. It is essential that the expert understands the difference between absolute scientific certainty which may be an unattainable ideal and the level of certainty required in most courts, typically termed "reasonable scientific certainty." An expert who is unprepared for cross-examination on this difference can be misled into making numerous damaging admissions about his or her lack of certainty that can sometimes be fatal to a case.

If you have a superior expert in any of the areas mentioned in Section A above, leading with this information at the outset of any direct examination is important and effective. If your expert has published numerous articles in peer-reviewed journals dealing with the subject matter of her testimony, this needs to be explained and highlighted. Although the term "peer-reviewed" is familiar to lawyers and judges, most jurors don't read technical journals and have no concept of the significance of the peer review process. Explaining it in detail is important, especially if the opposing side is not relying on peer-reviewed studies. By taking the time to explain why your expert is more qualified, and why the studies she relies upon are more reliable because they have been subjected to review by other scientists and deemed worthy of dissemination, you also set up the expectation in a juror's mind that the opposing expert should be as qualified to render the opposing opinion being offered and it too should be equally well supported. When the opposing expert is flawed, regardless of who testified first, explaining that your expert has actually arrived at his or her opinion through evaluation of scientific evidence leading to a conclusion, as opposed to being a hired gun who begins with a conclusions and then tries to find support for it, will go a long way even if the jury is unable to follow all of the technical data supporting your expert's opinion.

If there are studies supporting both the presence of a causal association and the absence of one, which is often the case, your epidemiology expert will have reasons for relying upon one group or the other that need to be explained. This is discussed at length in Chapter 5. Because nuances in epidemiological studies can be difficult for jurors to understand, it is important that your expert teach the jury in simplified terms why one group of studies is more reliable than the other, which will reinforce (in the case of the defense) or prepare the ground for (in the case of the plaintiff) cross-examination of the opposing expert on these studies. Again, it needs to be clear that the reliance on some studies more than others is based on sound scientific principles and not simply because the expert "likes" one set of results more than another.

One of the most difficult concepts to explain to jurors is statistical significance and the meaning of the p-value. This topic is discussed at length in Chapter 6 and should be reviewed in any case where one expert or the other has completely discounted studies on this basis or exaggerated the causal significance of a result that is statistically significant. An explanation of type one and type two errors, or the null hypothesis will also likely sail well over the heads of the average juror. Likewise, it is important to appreciate the difference between "statistical association" and "causal effect." Pains must be taken to explain these statistical concepts in the simplest possible terms. Covering these concepts thoroughly on direct examination to educate the judge during a *Daubert* hearing, or a jury during a trial, is paramount if your expert is going to rely on any studies the fall below the p-value magic line of statistical significance or if they find statistically significant associations that are not viewed as establishing the presence of a causal effect.

In Chapter 7 we discussed an objective methodology for evaluating the strengths and weaknesses of various studies to provide a better understanding of on which side the weight of the evidence rests as opposed to either lumping all studies together in a meta-analysis or cherry-picking the studies favoring a particular outcome. In cases where there are a number of conflicting studies, having the expert explain this methodology for ranking studies is important and will help establish the credibility of the expert in trying to reach a fair conclusion based on the weight of the evidence. This will tee up this issue for cross-examination, or provide a basis for your prior cross-examination of the opposing expert for his or her failure to follow such an objective methodology.

Published epidemiological studies are filled with complicated tables and graphs. While these are valuable to trained scientists as a way to display the data gathered and the analysis of these data, these graphics are typically not as useful to display to a jury. The evidence needs to be distilled to provide a clear takeaway message. Having the expert create simple tables and graphs that highlight the important findings of the studies and leave out what is unnecessary is usually far more effective. As long as the data on a summary chart or graph is derived from admissible evidence, there is usually no obstacle to using them to help get the main points across.

VII. CROSS-EXAMINATION

Cross-examining an expert in his or her own technical field is one of the most difficult and yet exhilarating parts of the job of any trial lawyer. With adequate preparation, frequently supplemented by assistance from your own expert, we are able to educate ourselves about a narrow scientific area to allow us to intelligently discuss a complex subject with someone who has spent a career in that field. The problem is that once we put in all of the work necessary to attain that level of knowledge, we tend to want to display it. In other words, we can wind up having a sophisticated technical conversation with the expert that is far beyond what a jury can comprehend. Accordingly, once we are thoroughly familiar with the studies relevant to the case, we can be in a position to support our arguments for or against a causal connection being present with a structured and methodical cross-examination of the opposing expert.

One universal truth about epidemiological studies is that no study is perfect and above criticism and without a weakness of some sort. Although weaknesses are on a continuum from fatal flaws that make a study virtually worthless to trivial uncertainties inherent in any study, an expert will always be able to point out some weakness about a given study. In preparing to cross-examine the opposing expert, counsel must be prepared to address the weaknesses the expert is likely to identify in discounting the results of each study your expert relied upon. Your expert will help in this preparation, and also in educating the judge and jury about the importance or unimportance of these weaknesses during his or her direct examination.

In Chapter 8 we discuss interpretation of negative studies, meaning studies that fail to show an association between the exposure and the outcome. There, we discuss negative studies conducted with solid methodologies, which are supportive of there being no association, vs. negative studies that have flawed methodologies where bias or lack of sufficient power is likely responsible for the result, limiting or even negating its value at assessing whether a causal connection is present. When negative studies are likely to be a feature of the opposing expert's presentation, it is important that you understand whether the negative study relied upon can be explained by any bias present, confounding or lack of sufficient power, and be prepared to confront the other expert on these factors on cross examination.

GLOSSARY OF EPIDEMIOLOGIC TERMS

In order to make epidemiologic research accessible to non-epidemiologists, which ultimately includes attorneys, judges, juries, and the public, it is important to minimize jargon and focus on making the concepts clear without resorting to esoteric terminology. Nonetheless, some basic terminology is helpful to orient the consumers of this research and enable them to follow and appreciate the logic used in interpreting epidemiologic evidence in the legal setting. There are a number of textbooks written for epidemiologists that can be consulted as needed but a brief enumeration of some of the key concepts is provided here, focusing on terms that have a different meaning than is used in normal communication. Many of the terms used have their familiar implications, e.g., measurement error, validity, precision, consistency, even if they are used in more formal ways by epidemiologists.

Attributable risk — the proportion of disease that can be attributed to a particular cause, calculated as 1-1/relative risk so for example, if the relative risk is 2.0, the attributable risk is 0.50 or 50%, meaning half the cases of disease are attributed to the exposure of concern.

Baseline risk — the frequency of disease in the absence of the exposure of concern, which is the denominator in a relative risk calculation or the hypothetical risk of disease among exposed persons had they not been exposed. The groups being compared based on higher or lower exposure should ideally have the same underlying baseline risk so that only the effects of exposure, if any, will cause them to differ.

Bias — the discrepancy between the measured association and the true causal effect, not an intentional misrepresentation or prejudice as in common usage. When we generate a measure of association but are really interested in the causal effect, bias can be defined as the deviation between what we measured and what are interested in.

Blinded study — blinding refers to who is aware of the study hypothesis, with a benefit from having study participants do not know whether they are receiving the active treatment or placebo (in a drug trial); single-blinded studies hide this information from study participants, and double-blinded studies hide this information from the researchers as well until the study is completed. It is used to make sure that any effects of the treatment are due to that treatment rather than expectations or beliefs about the effects of the treatment.

Bradford-Hill criteria—based on the work of Sir Austin Bradford-Hill, these considerations were offered to help determine when an established statistical association between exposure and disease is likely to be causal. Criteria include strength of the association, consistency, specificity, temporality, biological gradient, plausibility, coherence, experiment, and analogy.

Cause—attributes or exposures that directly influence the risk of developing disease. Although it is theoretical, if some of those who were exposed and developed disease had not been exposed, some of the cases of disease would not have occurred.

Case-control study—a design often used for studying relatively rare diseases in which the exposure history of those who developed the disease is compared to the exposure history of an appropriate comparison group, referred to as controls.

Cohort study—a design used in epidemiology that identifies groups with varying levels of exposure (the cohort) and monitors their disease risk over time, like an experiment but one in which the researcher observes rather than assigns exposure.

Confidence interval—in presenting the results of studies, the confidence interval is a way of expressing the spread or random error in the results, with small studies having wide confidence intervals to reflect greater uncertainty than larger studies. Typically, this is presented as a "95% confidence interval" which means that 95% of such intervals would contain the correct value.

Confounding—a source of bias in evaluating the causal effect of a specific exposure resulting from other causes of the health condition being correlated with the one of primary interest. When there is this mixing of effects from two or more exposures that tend to go together, there is a need to isolate the one of interest to avoid having it be confounded with the effect of other exposures.

Confounding by indication—in studies of the side effects of drugs used to treat disease, confounding by indication refers to the impact of the underlying disease for which the drugs were given which can be mistaken for the effect of the drugs themselves.

Controls—this is used in two different ways, which can be confusing. In experiments, it often refers to the group that lacks the exposure of interest, as in a drug trial comparing an active agent with no drug or placebo, which can be called the control condition. In case-control studies, it refers to the group without disease that is being compared to the group that has the health problem of interest. In the former instance, it means "lack of exposure" and in the latter it means "lack of disease."

Decision rules—guidelines for making judgments about the meaning of epidemiologic evidence may take the form of decision rules, which formalizes the inferences to be made based on specific types of evidence. While the formality of "if-then" algorithms may be appealing, it often results in simplistic reasoning that fails to consider the full range of relevant issues needed to make a judgment.

Dichotomy (artificial vs. pure)—division into two groups or categories. These can be pure dichotomies where the dividing line is clear (*e.g.*, "dead" or "alive"), based on convention (*e.g.*, "hypertensive" or "normotensive"), or arbitrary (*e.g.*, "sufficient" or "insufficient" evidence).

Differential and nondifferential misclassification—measurement errors can occur in both exposure and disease. An important determinant of the influence of those errors on study results is whether the pattern of error in one differs in relation to the other. If exposure measurement errors occur to the same extent both among those with and without disease, we refer to that as nondifferential exposure misclassification. Similarly, if disease measurement errors occur to the same extent among those who are and are not exposed, we refer to that as nondifferential disease misclassification. If the pattern of one differs in relation to the other, that is referred to as differential misclassification.

Dose-response relationship—if the disease risk increased across levels of exposure, a dose-response relationship is present, i.e., a graded response to varying levels of exposure.

Double-blinded study—studies in which the patients and the investigators are not aware of the exposure status of participants are referred to as double-blinded studies, often used in drug trials in which the active drug and placebo are indistinguishable. The purpose of this is to prevent patients or investigators from subtly influencing the outcome based on their expectations about the drug.

Ecologic study—study design that compares risks across groups of people rather than individuals, often using geographic units. Exposure and occurrence of disease are measured in multiple groups and then the relationship between exposure and disease is assessed by making comparisons of more highly exposed groups versus less exposed groups.

Evidence synthesis—integrating information from the full range of informative research to make a judgment about it is referred to as evidence synthesis. It can include generating pooled estimates of association through meta-analysis or other approaches to bringing the body of information together for evaluation.

False positive—an erroneous assignment in which the absence of some characteristic is incorrectly assigned as the present. This may apply to assignment of disease or exposure when the person is truly free of disease or exposure, or declaring an association or causal effect to be present when in truth it is not.

False negative—an erroneous assignment in which the presence of some characteristic is incorrectly assigned absent. This may apply to not assigning disease or exposure when the person truly has the disease or exposure, or failing to declare an association or causal effect to be present when in truth it is.

General causation—a causal relationship that is based on generalizable scientific information in which a set of individuals who have a certain attribute or exposure are believed to be at increased risk of developing disease because of that characteristic. This is distinguished from specific causation which refers to an individual whose health outcome is judged to have been affected by some attribute or exposure.

Generalizability—the ability to extend results from an individual study or set of studies to a broader population, generally used in judging whether apparent causal effects found in research would apply more generally.

Meta-analysis—integration of information from a set of studies on a given topic to calculate a pooled estimate of association, one that is a weighted average of the measure of association from each of the individual studies.

Negative study—a study that does not find an association between the exposure and disease of interest. This typically refers to null findings (no association) rather than an inverse association (a lower risk of disease among those who are exposed).

Null hypothesis—the hypothesis or assumption that there is no association between exposure and disease, i.e., that the relative risk is 1.0. This is often used as a benchmark to assess whether the study findings are sufficient to reject the null hypothesis and declare that an association is present.

Odds ratio—this measure of the association between an exposure and disease is one of a number of ratios that is calculated to describe how strongly associated they are. The literal meaning is the odds of disease among the exposed divided by the odds of disease among the unexposed or less exposed people.

P-value—the formal definition is the probability of having obtained results as or more extreme than those observed if the null hypothesis is true. As the measured association becomes larger and larger, the probability that such a finding would occur if there really is no association at all becomes increasingly remote. It does not provide a direct indicator of how likely it is that the results are due to chance or that there is any association present, but combines the size of the association and study size to help make those judgments.

Point estimate—in measuring an association, there is a calculated value that comes directly from the data, *e.g.*, a relative risk of 1.7, but this is often presented along with a confidence interval, *e.g.*, 1.2 to

2.3. Point estimate is contrasted with interval estimate which describes a range of possible values based on the confidence interval.

Pooled estimate—an aggregated estimate of the measure of the association, which is a weighted average of the measure of association across a series of individual studies.

Positive study—a study that finds an association between exposure and disease, contrasted with a negative study which does not find an association between exposure and disease.

Randomized clinical trial—a study design that randomly assigns individuals to receive some treatment (*e.g.*, drug) and others not to receive that treatment (*e.g.*, placebo or no medication). This is familiar as an experiment with exposure randomly assigned.

Recall bias—source of error in exposure determination that is based on participant recall, where those who have developed the health condition of interest more often report having been exposed than those who have not developed the health condition even if there is really no causal effect of exposure. This can result either from more complete recall among those with disease or overreporting among those with disease relative to those free of disease.

Replication—repeating a study using the same design and methods to determine whether it generates the same results as the original study.

Residual confounding—confounding that remains even after attempts are made to take account of the confounding factor and adjust for it. This can result from failure to measure the confounding factor accurately or completely so that the statistical adjustment is using an inaccurate measure of the confounder.

Risk difference—the result of subtracting the risk among the unexposed from the risk among the exposed. It is often contrasted with the risk ratio which is the risk among the exposed divided by the risk among the exposed. The risk difference accounts for how common the disease is overall whereas the risk ratio only considers how much the disease is multiplied as a result of exposure, not indicating how common or rare it is overall.

Risk ratio or relative risk—the risk among the exposed divided by the risk among the exposed. Sometimes contrasted with the risk difference, which accounts for how common the disease is overall, whereas the risk ratio only considers how much the disease is multiplied as a result of exposure, not indicating how common or rare it is overall.

Selection bias—a distortion in the measure of association that results from who provides data for the study. When those who enroll in the study or complete the study differ from those who were sought for the study, depending on the patterns of enrollees based on exposure and outcome, the resulting measure of association may be biased.

Sensitivity—the proportion of those who are positive that are identified correctly as falling into that category. This can be applied to identifying disease or exposure, or more broadly to identifying associations. Regardless, it addresses the completeness of identifying some attribute of interest, with the balance of those who truly have the attribute constituting false negatives.

Specific causation—this refers to a causal effect inferred for a specific individual, often contrasted with general causation which refers to a causal attribution for a group or population of interest.

Specificity—the proportion of those who are negative that are identified correctly as falling into that category. This can be applied to identifying disease or exposure, or more broadly to identifying associations. Regardless, it addresses the extent to which those who lack some attribute of interest are correctly identified as lacking that attribute, with the balance of those who truly have the attribute constituting false positives.

Statistical power—the statistical measure that estimates the proportion of times that a study would be capable of detecting an association as statistically significant if an association of a given magnitude is truly present.

Statistical significance—the probability of having obtained the results that were obtained if, in fact, there is truly no association present. Conventionally, the dividing line is often set at 0.05 or 5% so that when a study result has a 5% probability or less of having occurred randomly when there is truly no association present, it is declared to be statistically significant.

Underpowered study—a study that is too small to be capable of detecting an association of a given magnitude even if one is truly present.

Weighted average—a summary measure or average across a series of studies in which bigger studies are assigned a greater weight than smaller studies.