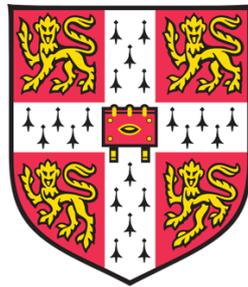


Predicting the Effectiveness of Medical Interventions



Adrian Dean Erasmus

Hughes Hall
University of Cambridge

This dissertation is submitted for the degree of Doctor of Philosophy

May 2021

Declaration

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared below and specified in the text. It is not substantially the same as any that I have submitted or is being concurrently submitted for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution. I further state that no substantial part of my dissertation has already been submitted, or is being concurrently submitted for any such degree, diploma, or other qualification at the University of Cambridge or any other University or similar institution.

- Part of Chapter 3 draws from published work, Erasmus, Holman, and Ioannidis (2020).
- Chapter 5 is based on a published paper, Erasmus, Brunet, and Fisher (2020).

This dissertation does not exceed the word limit of 80,000 words set by the Degree Committee of the Department of History and Philosophy of Science, University of Cambridge.

Predicting the Effectiveness of Medical Interventions

Adrian Erasmus

Abstract

This dissertation explores several conceptual and methodological features of medical science that influence our ability to accurately predict medical effectiveness. Making reliable predictions about the effectiveness of medical treatments is crucial to mitigating death and disease and improving individual and population health, yet generating such predictions is fraught with difficulties. Each chapter deals with a unique challenge to predictions of medical effectiveness.

In Chapter 1, I describe and analyze the principles underlying three prominent approaches to physical disease classification—the etiological, symptom-based, and pathophysiological models—and suggest a broadly pragmatic approach whereby appropriate classifications depend on the goal in question. In line with this, I argue that particular features of the pathophysiological model, such as its focus on disease mechanisms, make it most relevant for predicting medical effectiveness.

Chapter 2 explores the debate between those who argue that statistical evidence is sufficient for inferring medical effectiveness and those who argue that we require both statistical and mechanistic evidence. I focus on the question of how mechanistic and statistical evidence can be integrated. I highlight some of the challenges facing formal techniques, such as Bayesian networks, and use Toulmin’s model of argumentation to offer a complementary model of evidence amalgamation, which allows for the systematic integration of statistical and mechanistic evidence.

In Chapter 3, I focus on p-hacking, an application of analytic techniques that may lead to exaggerated experimental results. I use philosophical tools from decision theory to illustrate how severe the effects of p-hacking can be. While it is typically considered epistemically questionable and practically harmful, I appeal to the argument from inductive risk to defend the view that there are some contexts in which p-hacking may be warranted.

Chapter 4 draws attention to a particular set of biases plaguing medical research: Meta-biases. I argue that biases of this type, such as publication bias and sponsorship bias, lead to exaggerated clinical trial results. I then offer a framework, the bias dynamics model, that corrects for the influence of meta-biases on estimations of medical effectiveness.

In Chapter 5, I argue against the prominent view that AI models are not explainable by showing how four familiar accounts of scientific explanation can be applied to neural networks. The confusion about explaining AI models is due to the conflation of ‘explainability’, ‘understandability’, and ‘interpretability’. To remedy this, I offer a novel account of AI-interpretability, according to which an interpretation is something one does to an explanation with the explicit aim of producing another, more understandable, explanation.

Acknowledgements

I am immensely indebted to my supervisor, Jacob Stegenga, whose exceptionally insightful guidance and encouragement has been crucial to the writing of this dissertation. The intellectual rigor of his work is a constant inspiration, as is his generosity and ability to bring out the best in others.

My journey in philosophy began at the University of Johannesburg, where I was fortunate to have wonderful mentors. Thanks to Alex Broadbent, who cultivated my interest in philosophy of medicine. His guidance over the years has been invaluable. To Emma Ruttkamp-Bloem, for instilling in me a deep passion for this discipline. To past and present members of the UJ Department of Philosophy, particularly Thad Metz, H.P.P. Lötter, Catherine Botha, Johan Snyman, Raphael Winkler, Nicole Broadbent, Saana Jukola, Ben Smart, Oisín Keohane, and Neil Van Leeuwen, for their friendly and supportive counsel.

At the University of Cambridge, I have benefited greatly from being immersed in a vibrant interdisciplinary research environment, and from sharing my work with many outstanding scholars both in and beyond the Department of History and Philosophy of Science. Thanks, in particular, to my advisor, Anna Alexandrova, for providing exacting but constructive feedback on many aspects of this dissertation. For other valuable comments and discussions, I am grateful to Alexander Bird, Dan Steel, Tim Lewens, Jon Fuller, Rune Nytrup, Adrian Currie, Bennett Holman, John Ioannidis, and Eyal Fisher. I was lucky to be part of a superb group of graduate students. Thanks to Hamed Tabatabaei Ghomi, Cristian Larroulet Philippi, Zinhle Mncube, Victor Parchment, Elizabeth Seger, Adrià Segarra, Sophia Crüwell, and Oliver Holdsworth for their helpful feedback and camaraderie over the years.

Thank you to Lindsay and Nick Burton, Ariane Hanemaayer, and William Wong. The friendships I have formed with these brilliant individuals have contributed immensely to making this an enjoyable experience. I am particularly grateful to my fellow philosopher Tylore Brunet, with whom I have forged a fraternal bond, and from whom I have learnt much. His erudite input to my research is rivaled only by his kindness as a friend.

Thanks to the Oppenheimer Memorial Trust, the Cambridge Department of History and Philosophy of Science, and Hughes Hall College, Cambridge for helping to fund this research. Their financial support has been instrumental to the completion of this project.

My mother, Patricia Erasmus, has played a significant role in nurturing my academic path. I am eternally grateful to her for leading me to this point. Thanks to my father, James, brother, Denan, and all extended family, for providing a tremendous amount of love and encouragement over the years.

Finally, to Mandy Wigdorowitz, Polaris for a philosopher adrift, thank you for your immeasurable love, unending reassurance, and inexhaustible patience. Without you, this dissertation would not be what it is. Were the world founded on your capacity for care and compassion, the better we would all be for it.

Contents

Declaration	i
Abstract	iii
Acknowledgements	v
List of Figures	ix
List of Tables	ix
List of Abbreviations	x
Introduction	1
Chapter 1 A Pragmatic Approach to Disease Classification	8
1.1 Introduction	8
1.2 The Etiological Approach	10
1.2.1 The Monocausal Model of Disease	11
1.2.2 The Multifactorial Model	13
1.2.3 A Challenge for the Etiological Approach	15
1.3 The Symptom-Based Approach	17
1.4 The Pathophysiological Approach	21
1.5 Adopting a Broadly Pragmatic Approach	25
1.6 Conclusion	27
Chapter 2 Integrating Mechanistic Evidence for Inferring Medical Effectiveness	29
2.1 Introduction	29
2.2 What is Mechanistic Evidence?	31
2.2.1 Contrasting Statistical and Mechanistic Evidence	31
2.2.2 Mechanistic Evidence and Mechanistic Reasoning	33
2.3 Medical Evidence, Contextual Similarity, and Causal Consistency	34
2.3.1 Two Attitudes Toward Medical Evidence	35
2.3.2 Assumptions Regarding Contextual Similarity	37
2.3.3 Causal Consistency	40
2.4 The Toulmin Model: A Schema for Integrating Mechanistic Evidence with Statistical Evidence.	44
2.4.1 The Toulmin Model and Predicting Medical Effectiveness	44
2.4.2 Applying the Toulmin Model	47
2.6 Why the Toulmin Approach?	49
2.5 Conclusion	51
Chapter 3 P-hacking: Its Costs and When It Is Warranted	52
3.1 Introduction	52
3.2 What is P-hacking?	53
3.3 The Impact of P-hacking on Results	57
3.3.1 Why P-hacking Matters	57
3.3.2 A Case of P-hacking	59
3.3.3 The Prevalent Position and the Consequences of P-hacking	60
3.3.4 Immediate Challenges to the Prevalent Position	65

3.4	Warranting P-hacking	66
3.4.1	The Propriety of Analytic Choices	67
3.4.2	Warranting P-hacking using Non-Epistemic Judgments	68
3.5	Conclusion	70
Chapter 4 Curb Your Effectiveness: Correcting for Meta-Biases in Therapeutic Prediction		72
4.1	Introduction	72
4.2	What are Meta-Biases?	74
4.2.1	Methodological Bias and Meta-bias	75
4.2.2	How Meta-Biases Affect Estimations of Medical Effectiveness	80
4.3	Correcting for Meta-Biases: The Bias Dynamics Model	81
4.4	A Provisional Proposal for Determining Bias Coefficients	85
4.5	Conclusion	87
Chapter 5 Medical Artificial Intelligence: What is Interpretability?		88
5.1	Introduction	88
5.2	The Indefeasibility of Explanation	90
5.2.1	Four Kinds of Explanation	91
5.2.2	The Indefeasibility Thesis	94
5.3	The Explanation in the Machine: Medical AI Systems	98
5.3.1	Medical AI Systems Are Explainable	98
5.3.2	Separating Explanation and Understanding	102
5.4	Interpretability	103
5.4.1	What is Interpretation?	104
5.4.2	Total and Partial Interpretation	106
5.4.3	Local and Global Interpretation	108
5.4.4	Interpretation by Approximation or Isomorphism	110
5.5	Interpretation and Understanding	112
5.6	Conclusion	113
Conclusion		115
Appendix 1: The Expected Utility Principle		120
Appendix 2: Outcome Measures		122
References		123

List of Figures

Figure 2.1	An example of the simple Toulmin Model	45
Figure 2.2	An example of the extended Toulmin Model	45
Figure 2.3	The general form of the Toulmin Model	46
Figure 2.4	The Toulmin model for integrating mechanistic evidence with statistical evidence	46
Figure 2.5	Applying the Toulmin model for predicting medical effectiveness of efavirenz in Zimbabwe	49
Figure 5.1	The general structure of explanation	91
Figure 5.2	The structures of DN and IS explanation	92
Figure 5.3	The structures of CM and NM explanation	93
Figure 5.4	The structure of total interpretation	106
Figure 5.5	A partial interpretation wherein the explanandum remains the same in both the interpretans and interpretandum	107
Figure 5.6	A partial interpretation wherein the process of interpretation used results in a new process of explanation for an explanans and explanandum	108

List of Tables

Table 3.1.	60
Table 3.2.	65
Table 3.3.	65
Table A	120
Table B	121
Table C	122

List of Abbreviations

AD	Alzheimer's disease
ANN	Artificial neural network
ASA	American Statistical Association
CM	Causal mechanical
COPD	Chronic obstructive pulmonary disorder
DN	Deductive nomological
DSM	Diagnostic and Statistical Manual of Mental Disorders
EBM	Evidence-based medicine
ESS	Extrapolation by sufficient similarity
FDA	US Food and Drug Administration
FPRP	False-positive report probability
ICD	International Statistical Classification of Diseases
IHD	Ischemic heart disease
IS	Inductive statistical
MAIS	Medical artificial intelligence system
MDD	Major depressive disorder
ML	Machine learning
MTBC	Mycobacterium tuberculosis complex
NIHR	National Institute for Health Research
NM	New mechanist
RCT	Randomized control trial
RDFs	Researcher degrees of freedom
SEU	Simple extrapolation unless
WHO	World Health Organization

Introduction

Making reliable predictions about the effectiveness of medical interventions is essential to numerous goals in modern medicine. It plays a key role in diminishing death and disease and is vital to improving individual and population health. It is necessary for the development of successful healthcare policy and central to fostering public trust in the products, institutions, and practitioners of medical science. Yet, therapeutic prediction is fraught with difficulties. In this dissertation, I analyze several conceptual and methodological features of medicine that affect our ability to make reliable predictions about the effectiveness of medical interventions.

It is difficult to overstate the importance of understanding the elements involved in predicting medical effectiveness. Medicine is huge part of our personal and social lives. We are surrounded by guidelines, policies, and treatments aimed at mitigating suffering from disease and preventing unnecessary death in safe, cost-effective ways. Billions are invested into research on new therapeutics, and, in many cases, individuals commit thousands of dollars a year to health insurance for themselves and their families. In the midst of catastrophe, we turn to medicine to help reduce harms. At the center of all this is the notion that medicine gets it right, that we can trust what medical practitioners say will happen. If they get it wrong, people may die. It is thus not hyperbolic to say that understanding the complexities of predicting medical effectiveness is a matter of life or death.

While there is not enough space in this dissertation to cover all the features of medicine that contribute to predicting medical effectiveness, those I do explore fall under three relevant areas in medical epistemology. First, is the role of *evidence* in medical inference. Naturally, medical researchers and philosophers of science have emphasized the role of evidence in inferring the capacity of treatments to bring about their intended effects. In the realm of medicine, this emphasis on evidence has led to one of the most influential movements in modern science: Evidence-based medicine (EBM). At its inception, its early advocates deemed it “a new paradigm for medical practice” (Evidence-Based Medicine Working Group 1992). The aim of EBM is to utilize the best evidence for making inferences about medical interventions and diagnostic tests including pharmaceuticals, surgical procedures, and various screening programs. In line with this, EBM has developed a principled approach to evaluating the quality of various forms of medical evidence. This has resulted in the hierarchies of evidence we see today (see, for example, GRADE Working Group 2013). Proponents of EBM regard meta-analyses (along with systematic reviews) and randomized control trials (RCTs) as providing the best evidence for inferences of medical effectiveness. This is followed by observational studies like cohort studies and case-control trials. Typically relegated toward the bottom of evidence hierarchies are mechanistic evidence and expert opinion.

Not all medical researchers agree with the central tenets of the EBM movement. Some, including Donald Berwick (2005), an early advocate of EBM, believe that the dominance of epidemiological methods and statistical data in medicine has resulted in a significant disregard for

important types of scientific reasoning and evidence that, they argue, are necessary for the development of medical interventions. Often what is needed for such development, this argument goes, are small scale studies, lab benchwork, and good observational studies. Many philosophers of medicine have adopted a similar approach. Some have argued, in contrast to EBM, that mechanistic evidence is integral to making accurate predictions of medical effectiveness (Russo and Williamson 2007; Clarke et al. 2014; Parkkinen et al. 2018). Others have questioned the place of randomized control trials (Worrall 2002; Cartwright 2007) and meta-analyses (Stegenga 2011, Jukola 2015) at the top of evidence hierarchies, while some argue that the whole concept of evidence hierarchies is flawed (Upshur 2005; Solomon 2011, 2015; Stegenga 2014). Much of the work in this dissertation fits into debates about the role of medical evidence.

The second area I touch on is constituted by a pair of interrelated features relevant to predicting medical effectiveness: *Questionable research practices* and *research biases*. In recent years, questionable research practices, such as hypothesizing after results are known, data fishing, and p-hacking, have received just scrutiny. Many hold questionable research practices, at the very least, partially responsible for the replication crises facing psychology and medicine (Nosek et al. 2012; Open Science Collaboration 2015; Munafò et al. 2017; Christian 2017). There are other reasons to explore questionable research practices. If such practices are pervasive, then trust in medical science will be threatened (Anvari & Lakens 2018; Colquhoun 2017). Furthermore, the epistemic costs of questionable research practices may lead to harmful policies (Ioannidis and Trikalinos 2007; Simmons et al. 2011; Head et al. 2015;). Likewise, biases in research, such as outcome reporting bias and publication bias, have long been a topic of concern for medical researchers and, more recently, for philosophers of medicine. Such biases contribute to what some see as a general overestimation of medical effectiveness and underestimation of the harms caused by pharmaceuticals (Gøtzsche 2013; Stegenga 2018). Many of the arguments I present in this dissertation are influenced by, and extend on this previous work on questionable research practices and research biases.

Finally, a more recent area of concern is the reliance on *artificial intelligence* (AI) for aiding inference and decision making in medical contexts. There is a growing body of evidence showing that AI systems perform better than human experts in diagnosis, prognosis, and, importantly, making predictions about medical effectiveness (Fleming 2018; Rajpurkar et al. 2017; Tschandl et al. 2019). Many of these systems are being deployed in the clinic, across various subdomains of medicine. For example, algorithms are being used to generate breast cancer treatment recommendations (Somashekhhar et al. 2018) and make predictions regarding screening (McKinney et al. 2020). Many have argued that these successes are just the beginning, and that AI is set to become a staple in the process of medical inference (Wiens and Shenoy 2018). Of course, the implementation of AI in medicine is not without its problems. Some scholars have emphasized that the integration of AI in clinical decision making and prediction runs the risk of harmful outcomes. A notable example of this is the risk of algorithmic biases due to the methods used to train AI models seeping into treatment recommendations (Obermeyer et al. 2019). Of course, the question of how AI

models generate their predictions is important to those who use them and those who are meant to benefit from their use. This is a burgeoning field, and philosophers are well situated to cut through the conceptual weeds that obstruct discussions about the use of AI models in medicine. In this dissertation, I examine an important conceptual issue in medical AI, the concepts of *explainability*, *understandability*, and *interpretability*. While much more can be said about AI in medicine, my hope is that the arguments I make here will contribute to a deeper understanding of how algorithms can help make better predictions of medical effectiveness.

Given this wide array of themes, it is reasonable to ask whether this dissertation risks being spread too thin on the topic of predicting medical effectiveness. After all, it is common for projects of this kind to defend some overarching thesis in medical epistemology. However, by presenting five substantive chapters on somewhat distinct topics relevant to predicting medical effectiveness, I am able to focus in on more concrete issues that philosophers *and* medical researchers are concerned about. Moreover, I take it that this project, and other work like it, constitutes a logical next step to existing insightful monographs in the epistemology of medicine (Howick 2011b; Broadbent 2013; Solomon 2015; Stegenga 2018; Valles 2018).

With this in mind, the arguments presented in this dissertation should be valuable to three sets of readers. Since the research is primarily situated in philosophy of science, it is relevant to both professional philosophers and students of philosophy alike. The topics I explore are relevant to medical research and practice, and so my hope is that physicians and medical scientists, students, and organizations will, at the very least, find the work useful. Lastly, I wrote this dissertation with the underlying assumption that medicine ought to be practiced in service of the patient, and thus, the arguments I offer are pertinent to those who seek to benefit from therapeutic interventions. This dissertation is relevant to philosophers, medical practitioners, and patients.

Medical Effectiveness

Before outlining the structure of the dissertation, it would be useful to have a definition of its central concept: *Medical effectiveness*. To start, consider a widely accepted key distinction used throughout this dissertation, according to which the *efficacy* of an intervention is its capacity to bring about its intended outcome under ideal circumstances, whereas the *effectiveness* of an intervention is its capacity to bring about its intended outcome under normal circumstances (Cochrane 1972). Cartwright (2009; 2012) provides helpful clarification of this distinction. In the definition of efficacy what is meant “ideal circumstances” is usually *an experimental setting*. According to EBM, the ideal experimental setting is a well-conducted RCT. However, other experimental settings, such as those in various forms of observational trials, can provide somewhat ideal circumstances for claims of efficacy when RCTs are not viable. Cartwright further clarifies that claims which fall under the definition of efficacy are claims of “it works somewhere” (2012: 975). Effectiveness, on the other hand, includes claims of “it works generally” and claims that “it will work for us” (*ibid.*). The former, *it works generally*, is a claim about the general capacity of an intervention to bring about its intended effect in varied settings outside of the experimental context—there is no specific target context in

mind. The latter, *it will work for us*, is a claim about the capacity of an intervention to bring about its intended effect in a specific target context or population outside of the experimental setting. It is these kinds of claims that I associate with medical effectiveness in this dissertation. Thus, medical effectiveness is *an intervention's capacity to bring about its intended effect generally or in a target context*.

Stegenga makes a further distinction. When it comes to claims about medical effectiveness, he argues that we are often concerned with claims that it “works for me” (2018: 45); that is claims about whether an intervention will work for *a particular patient*. Such claims fall under what Fuller and Flores (2015) refer to as the “particularization” of clinical results. Claims of this kind are, as Stegenga rightfully argues, harder to warrant than *it works generally* since they rely on justifying not just that the intervention works generally, but also that the purported general capacity of the treatment applies to the particular patient in question. Following this line of reasoning, claims of *it works for me* are harder to justify than claims of *it will work for us*, which in turn are more difficult to justify than claims of *it works generally*. In the dissertation, I do not place explicit focus on the singularized sense of medical effectiveness, although I take it that claims of the kind *it works for me*, amount to extreme cases of claims that *it will work for us*; it is just that the target population in question is a population of one.

It should be noted that claims about medical effectiveness are often articulated quantitatively. In RCTs, study participants are randomly allocated into two groups, the *experimental group*, and the *control group*. Those in the experimental group receive the intervention being tested and those in the control group typically receive an inert intervention—a placebo—or, less commonly, no intervention. When a placebo is used, participants are usually blinded, meaning they are not informed about whether they are receiving the treatment or placebo. The efficacy of the intervention is then measured, commonly by researchers who are themselves blinded, by recording and comparing the outcomes for participants in each arm of the trial. Researchers quantify the outcomes of an intervention using an *outcome measure*. This is a formal expression of the relationship between the measured value of a property in the treatment group of the clinical trial, and the value of that property in the control group of the trial. Many different outcome measures are used in medical research (see Appendix 2), each of which generates a numerical result known as an *effect size*. The effect size is the measure of the difference that exposure to an intervention makes to a particular outcome in the study population of the trial. It is intended to illustrate the strength of purported causal relation between the treatment and an outcome in the experimental setting. Thus, the effect size in a study population constitutes a quantitative claim about the efficacy of the intervention in question. Once calculated, the effect size is typically used to infer the capacity of the intervention in question for some target population using a process of extrapolation.

Medical effectiveness is about more than (quantitative) claims regarding the general capacity of an intervention to bring about its intended effect. What, for instance, is meant by “bring about its intended effect”? Here, Stegenga (2018) provides a helpful argument. Stegenga argues for a hybrid

conception of disease, according to which a disease is a failure of particular physiological mechanisms to perform their particular functions at normal efficiency *and* a disvalued state. Since medical interventions target diseases, in order to count as effective, their intended effects must prevent or modulate either the physiological dysfunction—what Stegenga refers to as “the causal target of effectiveness” (2018: 29)—*or* the disvalued state—what Stegenga refer to as “the normative target of effectiveness” (2018: 30) of a particular disease. Integrating Stegenga’s insights into the previous characterization of medical effectiveness results in the following definition, which I will use throughout the dissertation:

Medical effectiveness is an intervention’s capacity to prevent or modulate the physiological dysfunction or disvalued state of a disease generally or in a target context.

Dissertation Structure

As I have mentioned, the topics covered in this dissertation only scratch the surface of the complex elements involved in predicting medical effectiveness. While I cannot explore every feature of therapeutic prediction here, I have aimed to cover topics of central concern to medical research practices. In what follows, I provide an outline of the dissertation.

Disease classification is imperative for our theoretical understanding of diseases. It plays a significant role in the practical and clinical goals of medicine, such as diagnosis, drug development, and prediction of therapeutic effectiveness. However, there is disagreement about the best approach to classifying diseases. In Chapter 1, “A Pragmatic Approach to Disease Classification”, I describe and analyze the principles underlying three prominent approaches to disease classification—the *etiological approach*, the *symptom-based approach*, and the *pathophysiological approach*. I then suggest a broadly pragmatic approach whereby classifications depend on the goal in question. In line with this, I argue that particular features of the pathophysiological model, such as its focus on disease mechanisms, make it most relevant for predicting medical effectiveness.

In discussions about the relationship between evidence and causal inference in medical research, two rival views have emerged. The first argues that good statistical evidence from randomized control trials or meta-analyses, is necessary and sufficient to warrant causal claims. The second suggests that both statistical and mechanistic evidence are necessary. This latter argument is often made on the grounds that prediction and extrapolation of medical effectiveness require knowing the underlying mechanisms of a purported causal relationship. However, this view raises a question: How should mechanistic and statistical evidence be integrated? In Chapter 2, “Integrating Mechanistic Evidence for Inferring Medical Effectiveness”, I clarify important aspects of these two competing views of medical evidence and offer an informal method for integrating mechanistic evidence with statistical evidence for predicting medical effectiveness. I differentiate between two senses of evidence integration, mathematical and systematic, and highlight some of the challenges facing formal techniques, such as Bayesian networks. I then use Toulmin’s model of argumentation

to offer a complementary model of evidence amalgamation, which allows for the systematic integration of mechanistic and statistical evidence.

Analytic choices in clinical studies can affect the reliability of predictions of medical effectiveness. *P-hacking*, the exploitation of statistical techniques such that trial results may be exaggerated, constitutes one such set of choices. In Chapter 3, “P-hacking: Its Costs and When It Is Warranted”, I provide much needed clarity on the concept of p-hacking and use philosophical tools from decision theory to articulate the prevalent view that p-hacking is epistemically and practically harmful.¹ Using this framework also allows me to illustrate the claims of vocal opponents to p-hacking and highlight the flaws in their arguments. While p-hacking is typically considered epistemically questionable and practically harmful, I appeal to lessons from the argument from inductive risk to defend the contrasting view that there are some contexts in which p-hacking may be warranted.

Medical research is fraught with numerous biases which affect our ability to reliably estimate medical effectiveness. In Chapter 4, “Curb Your Effectiveness: Correcting for Meta-Biases in Therapeutic Prediction”, I argue estimations of medical effectiveness often fail due to a particularly pernicious set of biases, which I refer to as *meta-biases*. Biases of this kind, such as publication bias and sponsorship bias, result in systemic overestimations of medical effectiveness and thus affect inferences about the effectiveness of treatments. Much research focuses on mitigating and preventing meta-biases, but these strategies often fail and when they are successful, their effects take time. To tackle these problems, I offer a framework, which I call the *bias dynamics model*, according to which estimates of medical effectiveness are corrected in line with up-to-date evidence on the prevalence and effects of meta-biases.

Artificial neural networks are fast becoming an integral part of healthcare and, importantly, are being used to aid in predictions of medical effectiveness. In Chapter 5, “Medical Artificial Intelligence: What is Interpretability?”, I argue against the prominent view that machine learning models are not *explainable* by showing how four familiar accounts of scientific explanation can be applied to artificial neural networks.² I suggest that the confusion about explaining AI models stems from an equivocation on the notions of *explainability*, *understandability*, and *interpretability*. To remedy this, I distinguish between these concepts in the context of AI, and provide a novel account of *AI-interpretability*, according to which an interpretation is something one does to an explanation with the explicit aim of producing another, more understandable, explanation.

My hope is that the arguments in this dissertation provide useful insights on these various challenges to predicting the effectiveness of medical interventions. While I do not defend an overall thesis, the work is important. There is a great deal at stake. Although it is unlikely that any single

¹ Part of this Chapter draws from work published on data-dredging in medical research (Erasmus, Holman, and Ioannidis 2020). Thanks are due to my co-authors on that piece, Bennett Holman and John P.A. Ioannidis.

² This chapter is based on a published journal article on the interpretability of artificial neural networks (Erasmus, Brunet, and Fisher 2020). Thanks are due to my co-authors on that paper, Tyler D.P. Brunet and Eyal Fisher.

project will solve all its problems, developing a better understanding of the concerns explored in the dissertation, however small, will contribute to improving predictions of medical effectiveness. What is more, this project will ideally open new avenues for philosophical exploration in the realm of therapeutic prediction. Some of these will be outlined in the dissertation's conclusion.

Chapter 1

A Pragmatic Approach to Disease

Classification

1.1 Introduction

Nosology is the part of medicine concerned with classifying diseases. One of its main aims is to improve communication among researchers, physicians, and health insurers. However, disease classification also contributes to numerous other goals in medicine, including therapeutic prediction. Given its significance in many areas, it is important to understand and evaluate the principles underlying the classification of diseases. In this Chapter, I clarify three prominent approaches to nosology and argue for a pragmatic approach to disease classification. In line with this, I suggest that we should use pathophysiological classifications in the context of predicting medical effectiveness.

Disease classification is vital to medical research. Its primary role is to inform standardized disease coding systems, such as the World Health Organization's (WHO) International Statistical Classification of Diseases (ICD), for use in epidemiological studies and medical reports on world health. However, the way in which medical researchers and practitioners interpret and pursue many other goals in medicine also depends, at least in part, on disease classification. Such classifications influence the formulation of diagnostic criteria and thus play a role in the process of diagnosis. For example, classifying osteoporosis as having a bone density that falls 2.5 standard deviations below mean levels in young healthy adults of the same sex (Lindsay and Cosman 2015) impacts the diagnostic criteria for the disease. Disease classifications can also provide important information about risk factors for various diseases. For instance, we might include information about the association between inadequate daily calcium intake and low bone density in our classification of osteoporosis. These are some practical aspects of disease classification that directly affect patients in the clinic.

There are less direct impacts too. Research practice and resource allocation, for instance, rely on classificatory practices. This is illustrated by recent proposals for the reclassification of Alzheimer disease (AD). AD is currently classified as a degenerative disease of the nervous system characterized by atypical amyloid plaque build-ups in the brain (Sennvik et al. 2000). However, there have been calls to reclassify AD as 'type 3 diabetes' (de la Monte and Wands 2008; Kandimalla et al. 2017). This proposal is based on growing evidence that the cognitive dysfunction associated with AD is caused by deficient utilization of glucose by the brain. Following this reasoning, AD is essentially a metabolic disease. Its reclassification would necessitate changes in the direction of AD research and

in intervention development strategies for the disease. Less emphasis would be placed on searching for ways to clear the disease's trademark amyloid plaque build-ups, and more focus and resources would be directed at understanding the metabolic disfunctions newly associated with the disease. This example also illustrates how the development of effective medical interventions depends on our understanding of individual diseases, which in turn is revealed in disease classifications. Given these impacts, it is crucial to ask how individual diseases should be classified.

In modern medicine, there are, broadly speaking, three prominent approaches to disease classification: The *etiological approach*, the *symptom-based approach*, and the *pathophysiological approach*. As a brief start, the etiological approach classifies diseases according to causes; the symptom-based approach classifies diseases according to symptoms or syndromes; and the pathophysiological approach classifies diseases according to underlying mechanisms. My first aim is a clarificatory one. I analyze these three approaches to disease classification, outline their distinctive principles, and touch on the advantages and disadvantages of each.

Despite its importance there is relatively little philosophical work on *physical* disease classification. Most scholarly work on nosology is limited to its role in psychiatry and the purported flaws of using a symptom-based model of classification for diagnosing and understanding mental disorders (Hempel 1965; Bolton 2012; Ghaemi 2012, 2013; Zachar and Kendler 2017). Other accounts examine the different models that fall under the etiological approach (Whitbeck 1977; Cooper 2002; Broadbent 2009, 2013; Smart 2014). Common to these accounts is the assumption that the etiological approach provides the most useful way to classify diseases. For example, when questioning psychiatry's systems of nosology, it is typically argued that an etiological model for classifying mental disorders would fare better than the symptom-based approach that has dominated the discipline since the publication of the DSM-III in 1980 (American Psychiatric Association 1980). And in discussions regarding etiological classifications, much emphasis is placed on whether diseases are best classified according to the *monocausal model of disease* or a *multifactorial model of disease* (see Section 1.2), without considering other strategies like a pathophysiological approach.

The etiological approach is privileged by several medical professionals as well. Scadding, for example, holds that “[w]hen the current state of knowledge permits more than one possible basis of definition of a disease, an aetiological basis will usually be more useful than a pathological, and a pathological than a syndromal” (1959, 323). In line with this, Snider makes the following claim:

These four levels indicate progressively decreasing knowledge of the disease and therefore decreasing priority as defining characteristics: etiology has the highest priority, altered structure or function, respectively, have intermediate priority, and clinical features have the lowest priority. (2003, 679)

To my knowledge, the only philosophical account to consider the pathophysiological approach is that of Fuller (2018b). He claims that all diseases are *in fact* defined according to their constitution. However, the reality is that physical diseases are often classified using other approaches. This could

be a result of epistemic gaps in our understanding of diseases. Perhaps we should read Fuller's claim that medicine in fact classifies according to the constitutive features of a disease as evidence that medicine has moved on from thinking etiological classifications are better than pathophysiological ones. If this is the case, then I disagree on pragmatic grounds. My second aim in this Chapter is to defend a broadly pragmatic approach to disease classification. That is, classificatory choices should depend on the goal being pursued. The picture of disease classification I suggest is as follows: The medical goal being pursued determines what kind of information about a disease is important, and this, in turn decides which classificatory approach should be used. This is because each model has particular advantages that are helpful in certain circumstances. To illustrate this point, I provide a brief argument for the use of the pathophysiological approach in the context of therapeutic prediction.

The Chapter proceeds as follows. In Section 1.2, I describe the etiological approach. Under this approach, diseases can be classified according to the monocausal model or the multifactorial model. Briefly put, the monocausal model of disease holds that, under the right circumstances, a particular disease is caused by a single necessary cause (Section 1.2.1). The multifactorial model, on the other hand, holds that disease causation is more complex and that a particular disease is the result of a constellation of causes (Section 1.2.2). Following this, in Section 1.3, I outline the symptom-based approach. Here, I discuss how physical diseases, like mental disorders, can be usefully classified by appealing to symptoms, despite the approach being largely rejected by many medical researchers. In Section 1.4, I discuss the pathophysiological approach. This approach uses mechanistic information to classify diseases. I outline two ways in which diseases can be classified under the pathophysiological approach: Either by appeal to abnormal anatomical *structure* or abnormal biomechanistic *function*. Then, in Section 1.5, I briefly outline the pragmatics of disease classification, touching on its use in therapeutic prediction. Finally, I conclude.

1.2 The Etiological Approach

A common way to classify diseases is by appealing to their etiology. The etiological approach to disease classification defines a disease in terms of the presence of its cause (or causes). In principle, this approach typically ignores other factors like pathogenesis, physiology, mechanisms, and symptoms. Put another way, etiological classifications of disease appeal to the *distal causes* of a disease, while disregarding its *proximal causes*. A distal cause of a disease is one that is extraneous to the constitutive mechanisms of the disease, and responsible for the manifestation and pathogenesis of that disease. Proximal causes, in contrast, are those that are constitutive of the underlying pathophysiology or mechanisms of the disease. Examples of distal causes include *Streptococcus pyogenes* for streptococcal pharyngitis, smoking for chronic obstructive pulmonary disorder (COPD), a mutation of the cystic fibrosis transmembrane conductance regulator (CFTR) gene for cystic fibrosis, or rheumatic fever for mitral stenosis.

The etiological approach seems reasonable given that it is commonly held that the best way to intervene on a disease is by modulating its causes (Scadding 1959; Snider 2003). However, the

attention it has received in philosophy of medicine suggests that it is problematic (Whitbeck 1977; Carter 2003; Cooper 2002; Broadbent 2009, 2013; Smart 2014; Fuller 2018b). In this section, I will provide analyses of two models of disease classification that fall under the etiological approach: The monocausal model and the multifactorial model. I will then briefly outline a criticism of the distinction between the monocausal and multifactorial models and offer a short argument for the usefulness of this distinction.

1.2.1 *The Monocausal Model of Disease*

According to the monocausal model, individual diseases are best understood as having a single necessary and universal cause. In the late Nineteenth Century and early Twentieth Century, breakthroughs of interventions that targeted specific pathogens became increasingly common. Such discoveries came to characterize modern Western medicine until the mid-Twentieth Century. This resulted in a paradigmatic shift from symptom-based classifications of diseases (see Section 1.3) to a causal model of classification. However, this also led to widespread acceptance of a relatively simplistic view of disease causation—that a disease has one cause. This view can be captured by the monocausal model of disease classification.

Broadbent (2013) offers a useful analysis of the monocausal model. On his account, the monocausal model is not simply the view that diseases have a single necessary and universal cause, but rather a single cause that satisfies the following two conditions:

- (i) The putative cause C_D is a cause of every case of the disease D .
- (ii) Under a specific set of circumstances insufficient to cause D , C_D causes D .

Broadbent refers to conditions (i) and (ii) as the *necessity condition* and *sufficiency requirement* respectively (2013, 150). The purpose of stipulating that C_D satisfy condition (i) is to ensure what Fuller refers to as “causal specificity” and “causal necessity” (2018b: 9). C_D is causally specific because it refers to one causal agent, not a disjunction of causes, and C_D is causally necessary because the disease will only occur if C_D occurs. Take *Streptococcus pyogenes* (*S. pyogenes*) for streptococcal pharyngitis; the only causal agent responsible for streptococcal pharyngitis is *S. pyogenes* and without it, streptococcal pharyngitis would not occur.

The key feature of condition (ii) is *circumstantial sufficiency* of C_D . Circumstantial sufficiency of C_D requires that a particular disjunction of circumstances, which are not jointly sufficient to cause D , must be present along with C_D in order for D to occur. According to Broadbent, the purpose of condition (ii) is to prevent the possibility of other causes satisfying the monocausal model of a particular disease. Without condition (ii), the argument goes, we could take any other cause, say lack of immunity to *S. pyogenes*, and characterize it as the necessary cause of streptococcal pharyngitis.

It is unclear that condition (ii) achieves this, however. Streptococcal pharyngitis can still, at least in principle, still be classified according to a necessary cause other than the presence of *Streptococcus pyogenes*. We could simply include the presence of *S. pyogenes* in our disjunction of circumstances

and characterize lack of immunity to *Streptococcus pyogenes* as the necessary cause of the disease. Of course, in practice this is not the case and the monocausal model is useful for classifying some diseases.

Classifying diseases monocausally is particularly useful for *infectious diseases*. These are diseases that are caused by pathogenic microorganisms such as bacteria, viruses, fungi, or parasites. For example, listeriosis might be classified as the infection caused by the presence of the bacteria *Listeria monocytogenes* in the body. In principle, classifying listeriosis in this way ignores the underlying mechanisms of the disease—that the bacterium enters the host’s phagocytic white blood cells, which leads to septicemia and permits access to the brain, causing meningitis, and allows for transplacental migration to the fetus in pregnant women.

According to the monocausal model, classifying diseases should adhere to the following *monocausal principle*, which states that a condition is a case of a particular disease if and only if the cause of the disease is the cause of the condition.

$$x \text{ is disease } D \text{ iff } C_D \text{ caused } x$$

Using listeriosis as an example, a condition (x) is a case of listeriosis (D) if and only if the presence of *Listeria monocytogenes* in the body (C_D) caused the condition.

The monocausal principle is simply Whitbeck’s (1977) *ontological conception* of disease classification. The ontological conception states that the identity of a disease is entirely determined by the disease-causing entity. On this view, classifying diseases entails merely identifying and distinguishing between different pathogens that could enter the body.

The monocausal principle can be a useful approach to classifying diseases. For one, it has some explanatory power. Classifying a disease by its cause partially explains why someone contracts the disease. For instance, categorizing tuberculosis as the disease caused by the presence of *Mycobacterium tuberculosis* (*M. tuberculosis*) partially explains why someone contracted the disease: *M. tuberculosis* entered the body which caused the pathogenesis of tuberculosis. An etiological disease category also partially explains why an intervention might be effective. For example, a monocausal classification partially explains why antibiotics are effective; they target the cause of the disease and thus eliminate the pathophysiological effects, and in turn the symptoms, of the disease.

Relatedly, striving to classify diseases using the monocausal principle may contribute to the development of effective interventions, such as vaccines, antivirals, and antibiotics, that target specific disease-causing pathogens. This is illustrated by commonly cited historical examples. Koch’s discovery of how to identify bacteria by staining along with his 1882 discovery of tubercle bacillus as the cause of tuberculosis laid the groundwork for Ehrlich to develop the concept of ‘magic bullet’ interventions. Magic bullets are interventions that target and eliminate the disease-causing pathogen, without damaging other cells in the organism. This ultimately led to the development of real-world ‘magic bullet’ interventions like the synthetic antibiotic arsphenamine (Salvarsan) for syphilis. This

paved the way for other advancements, such as Prontosil and, more famously, penicillin. Since these discoveries, numerous other antibiotics have been developed. Identifying a necessary and universal cause of a disease, and thus being able to classify the disease etiologically, is often crucial to the development of such treatments.

1.2.2 *The Multifactorial Model*

Of course, over the years epidemiologists and other medical researchers have determined that many diseases do not have a single cause but are, rather, multifactorial. Thus, classifying diseases according to multiple causes is another important model for disease classification that falls under the etiological approach. Some hold that all diseases, even those infectious diseases typically classified monocausally, are multifactorial (Krieger 1994; MacMahon, Pugh, and Ipsen 1960; Rothman 1976). Proponents of the multifactorial model consider diseases to be more complex than the monocausal model assumes and regard the causes of individual diseases as manifold. One way of thinking of this multifactorial model is as the simple claim that diseases have more than one cause.

On this view, we could refine the etiological approach so that it makes room for multiple causes for a disease. The monocausal principle could be reformulated as the following *naïve multifactorial principle*:

$$x \text{ is disease } D \text{ iff } C_1 \vee C_2 \vee \dots \vee C_n \text{ caused } x$$

Using the naïve multifactorial principle, a condition x is a case of osteoporosis (D) if the condition is caused by heavy drinking (C_1) or reduced amounts of testosterone (C_2) or any of the putative causes of the disease up to C_n .

However, this formulation would not be very useful since it entails that we think of diseases as being caused by any one of their putative causes. Furthermore, this would lead to confusion in classification since many diseases are caused by the same things. For example, a cause of small lung cancer is smoking, but the same can be said of osteoporosis. The multifactorial model is problematic in this sense since, on this naïve reading, it ambiguously characterizes the causes of a disease—is it that a disease could be caused by many singular causes, or by some set of causes, or by any one constellation of causes? Thus, this cannot be how medical scientists think of the multifactorial model.

The multifactorial model, as it is understood in medical science, holds that diseases are the result of complicated, congruent interactions of features including physiological, environmental, and lifestyle factors, and that individual cases of a particular disease are best described by appealing to these different causal risk factors for the disease. So, any particular disease might be characterized by appealing to different combinations of different lifestyle choices like smoking or inactivity, genetic features like having certain mutations, and environmental elements such as air pollution, which all increase the risk of contracting that disease.

Due to their causal complexity, when it comes to classifying *chronic diseases* etiologically the multifactorial model is typically used. Osteoporosis is commonly described using this causal risk factor understanding of the multifactorial model. Risk factors for osteoporosis include hyperthyroidism, reduced amounts of testosterone or estrogen, pituitary gland disorders, hereditary factors, heavy drinking and smoking, problems of malabsorption, long-term use of high-dose corticosteroids, to name just a few (Lindsay and Cosman 2015). Thus, an etiological approach to classifying osteoporosis would be to say that the disease is caused by any number of causal attributes associated with its risk factors.

Smart (2014) has proposed a more nuanced multifactorial model of disease classification that can make sense of this view. He claims that a disease can be classified according to “the disjunction of conjunctions of events jointly sufficient for contracting the disease” (Smart 2014, 257). Accordingly, the following *refined multifactorial principle* can be formulated:

$$x \text{ is disease } D \text{ iff } (C_1 \wedge C_2) \vee (C_1 \wedge C_3) \vee \dots \text{ caused } x$$

Smart claims that it is theoretically possible to classify diseases by outlining the finite possible causal configurations that could result in one contracting the disease.

What Smart is using in his analysis is essentially Mackie’s INUS account of causes (1965). On this view, a cause is an insufficient by necessary part of an unnecessary but sufficient—hence INUS—condition for an effect. That is to say, a cause of some effect is any member of a minimal set of sufficient conditions for that effect. In the case of disease, each member (C_n) of the finite possible causal configurations that result in one contracting the disease can be considered an INUS cause. The refined multifactorial principle above represents a special case where C_1 is present in every set of minimal sufficient conditions, but it should be noted that there need not always be a common cause of this kind. Applying the INUS account of causes to disease is helpful. For instance, it helps understand how smoking is a cause of osteoporosis: Smoking is an insufficient but necessary part of an unnecessary but sufficient condition for osteoporosis to occur. Similarly, testosterone deficiency is an insufficient but necessary part of another unnecessary but sufficient condition for the disease. Each disjunct in the refined multifactorial principle is a sufficient condition for a disease to occur, and each contains multiple INUS causes.

A problem for adopting the refined multifactorial principle is that most chronic diseases will, in a strict sense, be ill-defined. For instance, the causal risk factors for ischemic heart disease (*IHD*) are smoking, physical inactivity, obesity, hypertension, and excessive alcohol consumption (Antman and Loscalzo 2015), but none of these factors are strictly necessary for someone to develop the disease. Further, if all these causal risk factors were necessary, patients who suffer from IHD often differ in terms of the extent to which these factors are present and contribute to the development of the disease. That the set of causes for a disease could differ across individual cases of the disease brings the refined multifactorial principle into question as an appropriate way of categorizing individual diseases.

The above problem suggests that the number of different sufficient causal configurations for a disease are beyond human capacity to count. Smart (2014) responds to this objection claiming that similar inability is common in the sciences. He adopts an essentialist standpoint, according to which the aim of science is to discover essential properties of natural kinds. Not knowing these essential properties does not make essentialism unappealing, nor does it mean that the properties are unknowable and discovering more of the essential properties of natural kinds improves scientific understanding. Similarly, goes Smart's argument, discovering more causal configurations for a disease improves our understanding of the disease and allows us to manipulate these causes.

1.2.3 A Challenge for the Etiological Approach

A challenge facing the etiological approach stems from the monocausal-multifactorial distinction. Some philosophers have argued that the *prima facie* distinction between the monocausal and multifactorial models might not be as striking it seems (Broadbent 2009, 2013; Fuller 2018b). What does this mean for the etiological approach in general?

Diseases that are generally considered to be monocausal are the result of complex causal configurations that vary in individual cases and thus it is possible, in principle, to recharacterize them as multifactorial. This is particularly true for those cases where we do not have full knowledge of the specific set of circumstances in that make up part of condition (ii). For example, around 90% of patients who have *M. tuberculosis*, the bacterium responsible for tuberculosis, in their body do not develop any symptoms of the disease. This is known as latent tuberculosis. Active tuberculosis, on the other hand, is now understood as being caused by having the disease-causing bacterium in one's body along with some endogenous factors. But what those particular endogenous factors are is currently unknown. People with latent tuberculosis are often considered to be in a state of persistent immune response to *M. tuberculosis* and can develop active tuberculosis later in their lives if the presence of the bacteria is not treated. The manifestation of active tuberculosis in those with latent tuberculosis is normally put down to a weakening of the immune system, sometimes due to a secondary auto-immune deficiency such as HIV. However, the exact circumstances in each case are often unknown and cannot be explained by appeal to the monocausal classification. Thus, we can theoretically reclassify all purported monocausal diseases as multifactorial diseases.

The distinguishing feature of monocausal diseases, however, is that they are said to have a defining necessary cause, whereas multifactorial diseases do not. This amounts to monocausal diseases satisfying condition (i) but it is doubtful that many do. It is unclear whether even infectious diseases, considered exemplars of the monocausal model, satisfy causal specificity and causal necessity. Tuberculosis, for example, was once considered to only be caused by *M. tuberculosis*, but it is now understood that the disease can be caused by a member of the *M. tuberculosis* complex (MTBC), which includes *M. tuberculosis*, *M. africanum*, and *M. bovis*, among others.

Still, multifactorial diseases do not seem to have a defining necessary cause of the same sort as monocausal diseases. Given the successes of the monocausal model mentioned in Section 1.2.1, the fact that chronic diseases do not have a defining necessary cause might even be considered

unfortunate (Fuller 2018b). We know a vast amount about the causal risk factors of many chronic diseases. Smoking is considered a risk factor for IHD, yet we do not classify it using by this cause alone. Smoking as a cause for IHD is similar to *S. pyogenes* as a cause for streptococcal pharyngitis in that both increase the risk of contracting their respective diseases, but neither are sufficient without some other factors. While one may argue that casual risk factors merely contribute to disease, the same is true of the defining necessary causes of monocausal diseases, thus medical scientists could, at least in principle, use the monocausal model for chronic diseases.

In other words, in the same way that many infectious diseases have been reclassified in the past, we could simply reclassify chronic diseases monocausally (Broadbent 2013). Tuberculosis was once classified by its symptoms and was understood pathologically. It was considered to be caused by several infectious agents until Koch isolated *M. tuberculosis*. After this, tuberculosis was reclassified monocausally and, consequently, certain conditions that were once considered cases of the disease were excluded from the classification. The same can be done for chronic diseases. We could divide up the classification of IHD into several classifications such as ‘IHD caused by smoking’ and ‘IHD caused by high blood pressure’ and so on.

So, monocausal diseases can be classified multifactorially, and multifactorial diseases are not impervious to being classified monocausally. The distinction between the monocausal and multifactorial models thus seems not as stark as we might have thought, yet medical scientists tend not to use the monocausal model for chronic diseases nor the multifactorial model for infectious diseases. Why is that? In one sense, it is currently difficult to do so. The necessary causes of particular chronic diseases (whether these be some underlying genetic mutations or certain levels of exposure to environmental factors) are simply not as well understood as the necessary causes of particular infectious diseases are. In another sense however, chronic diseases are interpreted in a pragmatically different way to infectious diseases.

To understand what is meant by the pragmatic interpretations of infectious disease and chronic disease we need to consider how the etiological (distal causal) pictures of each are used. For infectious diseases, the etiological picture of there being a necessary universal cause for a disease is useful in the development of interventions that target those distal causes. Adopting the view that infectious diseases are monocausal aligns well with our historical capacity to develop such treatments, such as antibiotics and antivirals—it has worked. Adopting a multifactorial etiological picture of infectious diseases would arguably lead to failed research into less practical causal targets. Of course, vaccinations are aimed at developing immunity to particular pathogens, which may be interpreted as targeting a different cause of infection, namely a lack of immunity to the pathogen in question. However, even the development of vaccines requires that we accept the monocausal interpretation, since such research relies on identifying the pathogen we wish to prevent from infecting people. There is more success to be had by adopting a monocausal etiological picture of infectious diseases.

Chronic diseases, on the other hand, are interpreted multifactorially because their causes are not as explicitly linked to the pathogenesis of disease as those of infectious diseases. The distal causes of

chronic diseases are commonly viewed as risk factors, and not as strict causes. This is because while we may know that certain causal features are risk factors for a disease—features that contribute to an increased risk of developing the disease—we often do not know the extent to which those features contribute to the disease. Thus, we cannot be as clear about when a certain risk factor is necessary for a chronic ailment as we can for a particular pathogen and an infectious disease. Relatedly, we often cannot intervene on such risk factors in the same way as we can on the causes of infectious diseases. It is not the case that we can intervene on many distal causes for chronic diseases, such as smoking, heavy drinking, and obesity, in the same way that we can for those of infectious diseases. We might argue for reducing one’s exposure to such factors, but if one is exposed, (a) there is no certainty that they will develop the disease, and (b) if the disease did manifest we would often be unable to give them a drug and remove that exposure in the same way as we would a pathogen. We interpret chronic ailments using a multifactorial etiological model because useful necessary universal causes, *sensu* the monocausal model’s interpretation of a necessary universal cause, do not exist or are not known for such diseases.

This indicates that pragmatic considerations relating to identifying factors that can be usefully intervened upon, rather than purely logical considerations relating to necessary causes, explain why infectious diseases are typically classified monocausally and chronic diseases multifactorially. Thus, an etiological approach to disease classification is pluralist in the sense that, depending on the pragmatic understanding of the disease in question, classifications can be monocausal or multifactorial. As I argue later in the Chapter, pragmatic considerations are inseparable from our disease classificatory practices.

1.3 The Symptom-Based Approach

So far, we have established that the etiological approach can use either the monocausal model or multifactorial model to classify diseases by their cause(s). Another approach to disease classification involves appeal to the observable clinical picture that a patient presents; that is, the symptoms or syndrome of the disease. For example, classified symptomatically, mitral stenosis is characterized by dizziness, shortness of breath, fatigue, heart palpitations, and chest pain (O’Gara and Loscalzo 2015). This is the symptom-based approach to disease classification. Since there is little contemporary philosophical work on the symptom-based classification of physical disease, I discuss the symptom-based model of classification with reference to its widespread use in psychology and psychiatry. It should be noted that there are examples of physical diseases that are classified according to symptoms. When I analyze the approach through the lens of psychiatry, I do so to appropriately clarify certain aspects of the model, such as its advantages and the challenges it faces.

In this section, I also identify a major problem facing a symptom-based classification: The problem of underdetermination of treatment strategies due to shared symptoms. I argue that to avoid this problem, a symptom-based model of classification should be such that it reveals the underlying structure or function of the disease. This however, I argue, entails that we are already able

to classify these diseases according to their pathology or pathophysiology. This means that there are some cases where symptom-based classifications are redundant.

The symptom-based approach is common practice in psychiatry, particularly after the APA adopted an *atheoretical* approach to disease with the publication of the DSM-III in 1980. In the ICD-10, unless the etiology is known, mental disorders are also classified by their symptoms. An exemplar of a nosological system that predominantly uses a symptom-based approach is the DSM-5, the current version of the APA's Diagnostic and Statistical Manual of Mental Disorders. Here is an example of a disorder which is classified symptomatically in the DSM-5. Major depressive disorder (MDD) is a state in which at least five of the following symptoms are present during the same two-week period, with at least one of the five symptoms being either symptom (1) or symptom (2):

- (1) Depressed mood on most days for most of the day, indicated either by subjective report or by observation of others.
- (2) Notably reduced interest or enjoyment in all, or almost all activities, most of the day, almost every day.
- (3) Significant weight loss without diet or weight gain, or an increased or decreased appetite.
- (4) Insomnia or hypersomnia almost every day
- (5) Psychomotor agitation or retardation most days (observed not just subjectively reported)
- (6) Feeling fatigued or loss of energy most days.
- (7) Feelings of worthlessness or excessive unwarranted guilt most days.
- (8) Reduced ability to concentrate, or indecisiveness most days.
- (9) Recurring thoughts of death, or suicide with or without specific plan, or attempted suicide.

Two caveats for the diagnosis of MDD include that the symptoms must cause significant social or occupational distress or impairment, and that the condition must not be the result of taking any substance or other medication (American Psychiatric Association 2013).

With this model of classification, the following *symptom-based principle*, which states that a condition is a case of a particular disease if and only if the set of symptoms of the condition are the same as a subset of symptoms of the disease, can be formulated:

$$x \text{ is disease } D \text{ iff } G_x = G_{D1} \vee G_{D2} \vee \dots \vee G_{Dn}$$

A condition (x) is MDD (D) if and only if the condition is a set of symptoms (G_x) that are the same as at least five of symptoms (1) – (9) and at least one of those symptoms is symptom (1) or (2) (G_{Dn}).

Physical diseases can also be categorized by their symptoms. This is most common for those diseases for which only the symptoms are known. A good example is *nodding disease*—a mentally

and physically disabling disease only affecting children, which is confined to small regions of South Sudan. The symptoms of nodding disease include a total stunting of growth affecting not just the body, but the also the brain. It is named for its characteristic nodding seizures which occur when affected children begin eating or, in some cases, when they feel cold. Nodding disease is described according to its symptoms because its cause is currently unknown (Centers for Disease Control and Prevention 2019). Another example of a physical disease that can only be categorized by its symptoms is acute flaccid myelitis (AFM), a neurologic illness which, like nodding disease, affects children. AFM presents as localized weakness of the limbs, but its cause is unknown (Hopkins 2017).

A symptom-based approach to disease classification is useful in some cases. For one, developing a nosological system based on symptoms would mitigate the effects of theoretical disagreement that is sometimes present in medicine. This is one of the reasons for the APA's "atheoretical" approach to the causes of mental disorders since 1980 (American Psychiatric Association 1980, 1987, 1994, 2000, 2013). There is often disagreement in the mental healthcare profession about the causes and appropriate treatment of mental disorders. By adopting an atheoretical approach, the APA aims to reduce the alienation of any particular constituencies in the profession and compels mental health professionals to diagnosed disorders based on agreed upon symptoms. Although there is a significant historical literature on this debate, the focus of this Chapter is on physical disease—this short discussion of psychiatric nosology is useful insofar as it illustrates a possible similar advantage that might be applicable when categorizing physical diseases where consensus on their causes has not been reached.

Another related use of adopting a symptom-based approach is its potential benefits to the process of diagnosis. Nosology is often equated with diagnosis because the ways in which diseases are classified are often used to make inferences about what disease a patient may have. However, the two are distinct in an important way: Diagnosis is inferential whereas nosology is descriptive. That is to say, diagnosis is an iterative process of inferring what disease a patient has based on information gathered about particular characteristics of the patient. This includes information gathered from the patient's clinical history and from a patient interview, which includes questions about the patient's more recent history and their perceived symptoms. This is followed by a physical exam to gather other information that might contribute to formulating a working diagnosis. A working diagnosis is then formed based on the physician's interpretation of the information, followed by diagnostic tests to either confirm or disconfirm the diagnosis. If the diagnostic tests confirm the working diagnosis, an intervention strategy is prescribed. If the diagnostic tests disconfirm the working diagnosis, it is refined or reworked based on the new information gathered through diagnostic tests.

Nosology, on the other hand, is concerned with typologizing diseases. Medical researchers pick out defining characteristics of diseases and formulate classifications based on these characteristics. Of course, the logic of nosology is still fundamental to diagnosis since inferring diagnoses is only possible using an underlying nosological system (Armstrong 2011). When a working diagnosis is formed, the information gathered is constituted, in part, by reported and observed symptoms, so

perhaps having symptom-based categories for diseases would provide doctors with a nosological system that aids in this process. Take MDD for example: once it is established that a patient has at least five of the disorder's symptoms, with at least one of the five symptoms being either symptom (1) or (2), then they are generally diagnosed with MDD. Disease classifications provide definitions which determine diagnostic criteria which in turn inform physicians whether their diagnoses are likely accurate or not.

The benefit of having a set of criteria against which to measure a diagnosis, however, is shared by all three models of disease classification. Furthermore, it seems that a symptom-based model of nosology is not better than other approaches unless it includes some theoretical aspects. This is because of the following particularly serious challenge facing any symptom-based nosological system: In some cases, symptoms underdetermine decisions regarding appropriate disease intervention strategies. This is a common issue for physicians. To illustrate, consider the following scenario. A patient is brought into an emergency room. Her symptoms include trouble walking and speaking, partial paralysis of the face, trouble seeing in her left eye, and a headache. Based on these symptoms, her clinical history, and an interview with relatives who brought her in, the patient is diagnosed with stroke. Considering that time is of the essence, emergency measures must be taken. At this point, doctors in this emergency room have two options available to them: either they administer an intravenous injection of a tissue plasminogen activator (TPA), or they give the patient an antihypertensive drug.

Given that a TPA injection targets the cause of ischemic strokes but is contraindicated in hemorrhagic strokes, and an antihypertensive drug targets the cause of hemorrhagic strokes but is contraindicated in ischemic strokes, how should the decision regarding which treatment strategy to use be made? The reported and observed symptoms do little to help with the decision. Normally, the physician orders a CT scan, the results of which would show if there were bleeding in or around the brain. If the results show that there is no bleeding, the doctors can infer that the stroke is ischemic. If the results show that there is bleeding, then the doctors can infer that the stroke is hemorrhagic. This then allows the doctors to decide on which intervention strategy is best.

There is a crucial difference between the reported and observed symptoms and the one revealed (or shown to be missing) by the CT scan. The latter symptom is one that reveals important information about the underlying pathology (and pathophysiology) of the disease in question, whereas the former set does not. Thus, only symptoms which reveal the underlying mechanisms and causes of a disease are good for determining an appropriate treatment strategy.

While a symptom-based model of classification is problematic, knowing the symptoms of diseases can contribute to other goals in medicine, such as making some decisions regarding the direction of research into interventions. We might observe a symptom in one patient that was present in another and infer that an intervention that proved effective for the latter would be effective for the former. Analgesics, for instance, are effective interventions for the pain associated with many diseases. Of course, such inferences do not necessarily establish that the intervention will

target the causes of a disease, but they do provide reason to expect that the intervention will alleviate the symptoms of some condition. Such interventions lean further toward disease management rather than disease elimination. Given this, it would be hasty to completely discount a symptom-based model of nosology, especially in cases where the causes of a disease are unknown. And in fact, a similar symptom-based inference is made regarding nodding disease, where anticonvulsants are often used to help mitigate the seizures caused by the disease.

The symptom-based approach to disease classification, thus, has some important uses. It is a good starting point for understanding diseases and thus helps in reasoning about treatments. It can be useful for developing new disease categories when there are epistemic gaps regarding other properties of a disease. It is useful for the diagnostic process. Finally, as long as we are clear about its limits, the symptom-based approach can be useful for decisions regarding treatments. Symptoms used for treatment decisions should reveal important information about the underlying pathophysiology of a disease. In the next section, I discuss the pathophysiological approach to disease classification.

1.4 The Pathophysiological Approach

The last prominent approach to disease classification is the pathophysiological approach. This approach aims to classify disease according to what they are, their *constitution*. The kinds of features that constitute a disease are largely determined by the contemporary naturalist conception of disease. According to this approach a disease is abnormal functioning of a part or process in the body (Boorse 1977). The pathophysiological approach, then, aims to classify diseases by appeal to the parts and processes of the body that are dysfunctional. This notion of dysfunctional parts and processes can be directly linked to speak of mechanisms in biology and medicine. Machamer et al. (2000) define a mechanism as entities (or parts) and activities (or processes) and their organization (or their normal functioning). Thus, the pathophysiological approach classifies diseases according to mechanisms.

In this section, I examine the pathophysiological approach. I touch on the distinction between mechanisms in terms of pathology (structure) and pathophysiology (function) noting that since abnormal structure is implied by abnormal functioning, the distinction is a matter of degree. What this means is that functional abnormalities are caused by structural abnormalities and thus when we speak of a functional abnormality we are implying that a structural abnormality is present. I conclude that, since the practical effects of using pathological classifications are indistinct from those resulting from using pathophysiological classifications, the latter model of classification is sufficient for achieving the goals for which the former might be used.

As I mention above, classifying diseases by their pathophysiology involves defining a disease by what it is in the mechanistic sense. In structural biology, the relation between *structure* and *function* is emphasized. Structure refers to the form and makeup of biological elements like cells, proteins, and DNA. For example, bacteria are prokaryotic cells meaning they lack a true nucleus. Function, on the other hand, refers to the mechanistic processes performed by the structural elements. For

example, a functional description of bacteria is that they reproduce by binary fission. Biology commonly classifies living things by both structural and functional similarities.

Similarly, in the biomedical sciences, diseases can be classified by structure or function. The pathology of a disease refers to abnormalities in the body's structures, whereas the pathophysiology of a disease refers to abnormalities in the body's functions. One might assume that we can distinguish between pathological classifications of disease and pathophysiological classifications. On this view, the pathological classification of a disease involves describing structural abnormalities of the anatomical site in question, while a pathophysiological classification involves describing abnormal physiological function. On closer analysis however, we see that when it comes to disease classification the distinction between pathological and pathophysiological is just a matter of degree of detail and a pathophysiological model of classification is sufficient for a multitude of biomedical goals.

To see how the distinction between pathological and pathophysiological classification dissolves, I will describe classificatory principles for each. I will use the example of mitral stenosis to illustrate my argument. A pathological classification of mitral stenosis is *a narrowed apex of the funnel-shaped mitral valve with a shrinkage of the valve leaflets and presence of fibrous tissue and/or calcific deposits* (O'Gara and Loscalzo 2015). A corresponding *pathological principle* for disease classification, which states that a condition is a particular disease if, and only if, that condition is a certain pathological structure, can be formulated as follows:

$$x \text{ is a case of disease } D \text{ iff } S_x = S_D$$

Using mitral stenosis: A condition (x) is a case of mitral stenosis (D) if, and only if, the mechanistic structure of the condition S_x is a narrowed apex of the funnel-shaped mitral valve with a shrinkage of the valve leaflets and presence of fibrous tissue and/or calcific deposits (S_D). In the pathological principle, S_D refers to specific structural abnormalities. Importantly, S_D can be finer- or coarser-grained in its description of the structure, which entails describing the pathology of the disease at different physical scales. For example, a coarser-grained description of S_D for mitral stenosis would be *a narrowed mitral valve orifice*.

The pathophysiological classification of mitral stenosis could be *an abnormally elevated left atrioventricular pressure gradient due to significant obstruction of the mitral valve orifice area*. A corresponding *pathophysiological principle* for disease classification, which states that a condition is a particular disease if and only if the underlying mechanisms of the condition are the same as the underlying mechanisms of the disease, is the following:

$$x \text{ is a case of disease } D \text{ iff } F_x = F_D$$

Using mitral stenosis once more: A condition (x) is a case of mitral stenosis (D) if and only if, the underlying mechanisms of the condition (F_x) are an abnormally elevated left atrioventricular pressure gradient due to significant obstruction of the mitral valve orifice area (F_D). In the pathophysiological principle, F_D refers to specific functional abnormalities.

The functional abnormalities appealed to in F_D can include more, or less specificity and may describe fine- or coarse-grained descriptions of the mechanisms involved. A less specific mechanistic descriptions of mitral stenosis is *an abnormally elevated left atrioventricular pressure gradient due to significant obstruction of the mitral valve orifice area*. A more specific mechanistic definition of the disease is *an elevated left atrioventricular pressure of $\sim 25\text{mmHg}$ due to a mitral valve orifice area of $<1.5\text{cm}^2$* (O’Gara and Loscalzo 2015).

A coarse-grained mechanistic definition of mitral stenosis is *a reduction of blood flow in the heart*, and a fine-grained mechanistic definition is *a reduction of normal blood flow from the left atrium to the left ventricle due to a narrowing of the mitral valve orifice area*. A finer-grained mechanistic definition of the disease is *an elevated left atrioventricular pressure of $\sim 25\text{mmHg}$ due to a mitral valve orifice area of $<1.5\text{cm}^2$* . And finer still: *An elevated left atrioventricular pressure of $\sim 25\text{mmHg}$ due to a mitral valve orifice area of $<1.5\text{cm}^2$ which is the result of an abnormally elevated left atrioventricular pressure gradient due to significant obstruction of the mitral valve orifice area*.

What is noteworthy here, is that with more specificity and finer-grained descriptions of the *pathophysiology*, we can include descriptions of the *pathology* of the disease in question. On this account, pathological classifications are constrained by the structural descriptions in S_D . That is, if we were to attach anything other than structural descriptions to S_D , the pathological principle would no longer produce purely pathological classifications. Conversely, a classification made using the pathophysiological principle may include structural descriptions to provide a more detailed description of the mechanisms at work in the disease and remain a pathophysiological classification. Thus, to distinguish between the pathological and pathophysiological classifications is, at best, a matter of degree, or, at worst, incorrect. We should, I propose, opt for the pathophysiological principle over the pathological principle.

Let us analyze how the pathophysiological approach fares against the challenge faced by the symptom-based model where the elements appealed to in a disease classification underdetermine the appropriate treatment strategy. The pathophysiological principle entails that a condition (x) is a stroke (D) if and only if the underlying mechanisms of the condition (F_x) are such that brain tissue is being damaged due to insufficient blood flow to the brain (F_D). Stroke is a broad category of a disease which obviously lacks the detail required in some parts of the biomedical sciences. However, as we have just seen, one advantage of a pathophysiological approach to disease classification is that it can provide granular descriptions of diseases. One might, for instance, include the categories of ‘ischemic stroke’ and ‘hemorrhagic stroke’ under the broader category of stroke. Each would include finer grained descriptions of the underlying mechanisms of these types of stroke, ones that include the pathophysiology-revealing symptoms to differentiate between different types of stroke. For instance, an ischemic stroke is a condition where brain tissue damage is the result of insufficient blood flow to a part of the brain *caused by a blockage in blood supply*. A hemorrhagic stroke, on the other hand, is a condition where brain tissue damage is the result of insufficient blood flow to a part of the brain *caused by bleeding in and around the brain*. The usefulness of these finer-grained

classifications manifests in their contribution to decisions regarding appropriate intervention for a condition, discussed in Section 1.3 above.

Discovering the underlying mechanisms of diseases has led to numerous effective medical treatments, and thus using a pathophysiological model of nosology might be beneficial for further development. For instance, in 1920 Frederick Banting and Charles Best discovered that type 1 diabetes is the autoimmune destruction of insulin-producing beta cell population of the pancreas, which prevents the pancreas from producing insulin, which leads to high glucose levels. From this they inferred that insulin therapy might mitigate the effects of type 1 diabetes. This led to one of the most effective medical interventions we have today. Although autoimmune damage to the pancreas still occurs, insulin therapy is effective because it modulates the body's inability to produce insulin, thus providing a way to manage the effects of the disease.

At first glance, a pathophysiological model of nosology seems to favor a multifactorial model of disease causation. Chronic diseases like type 1 diabetes, osteoporosis, and IHD are easily categorized using a pathophysiological model, and furthermore, are classified at broader- or finer-grains depending on the goal in question. For instance, IHD could be defined at a fine grain by referring to the mechanisms involved in the process. Here, IHD may be classified as inadequate supply of blood and oxygen to a portion of the myocardium “due to a combination of *fixed vessel narrowing and abnormal vascular tone* as a result of *atherosclerosis and endothelial dysfunction*” (Yelle 2012; my emphasis).

Given that it is well suited for defining chronic diseases, what use is the pathophysiological approach for infectious diseases? Infectious diseases can be described using the pathophysiological principle too. Take pulmonary tuberculosis: for a condition (x), if the underlying mechanisms of the condition (F_x) are such that inhaled bacilli from the MTBC reach the alveoli of the lungs where macrophages phagocytose the bacilli, which leads to multiplication of the bacilli and further infection and extracellular infection gradually forming granuloma which eventually necrotize leading to blood and sputum in the lungs (F_D), then the condition is a case of pulmonary tuberculosis (D) (Raviglione 2015). Classification according to pathophysiology in this way is possible for all infectious diseases.

If diseases which can be described according to their causes can also be described according to their underlying mechanisms, we might have reason to question the supposedly stark distinction between the etiological and pathophysiological systems of nosology. However, while the etiological model is restricted to distal causes, the pathophysiological model I am proposing, in virtue of describing diseases mechanistically, includes proximal and, at times, distal causes in its disease definitions. Take the following pathophysiological description of type 1 diabetes. The underlying mechanisms of type 1 diabetes are such that the immune system mistakenly destroys the insulin producing beta-cell population of the pancreas which prevents the pancreas from producing insulin which in turn leads to high glucose levels in the blood. This pathophysiological explanation of type 1 diabetes has causal components—the autoimmune destruction of beta cells causes insulin

deficiency, which in turn causes high glucose levels in the blood. Another good example is listeriosis. The pathophysiological mechanisms of listeriosis entail that *Listeria monocytogenes* bacteria present in the body enter phagocytic white blood cells, which causes septicemia, and which allows the bacteria access to the brain, causing meningitis. This pathophysiological explanation of listeriosis contains causal information about the distal cause and the proximal causes of the disease.

Because of the option to include proximal and/or distal causes in the definitions of diseases when using the pathophysiological principle, the pathophysiological approach is appropriate for informing the development of medical interventions. Interventions are often developed based on mechanistic evidence; thus, it is reasonable to expect the pathophysiological model to contribute greatly to this goal.

This means that medical researchers and practitioners have various options open to them when using the pathophysiological approach to disease classification. Of course, if there is little understanding of the pathology or pathophysiology of a disease, a symptom-based model will be useful for directing research to further develop our mechanistic knowledge about the disease. If, however, our pathophysiological understanding of a disease is good, then we are faced with choosing between structural or functional classifications. Sometimes, good symptom-based classifications are sufficient to serve a given aim, such as a physician informing a patient of how they inferred a particular diagnosis. Other times, a pathophysiological classification is more suitable, such as when the goal is intervention development. Moreover, because of the possibility of providing more specific and/or finer-grained (or less specific and/or coarser grained) classifications of diseases, the pathophysiological approach can be further manipulated to serve very specific goals in medicine.

1.5 Adopting a Broadly Pragmatic Approach

What is clear from the discussions in the previous sections is that disease classifications come in many forms and that medical researchers and physicians have multiple options to choose from when deciding how to classify diseases. A primary advantage of this is that medicine can (and should) adopt a broadly pragmatic approach when it comes to disease classification. The disease-related medical goal being pursued should determine which classification of the disease in question is used. Given that each approach outlined above has its own advantages, there are different aspects of medicine that could benefit most from using a particular approach. This is not exactly a novel idea—medicine has adopted a broadly pluralist approach to classification. Diagnosis, construed as a primary aim of medicine, has benefited from the use of symptom-based approaches since the dawn of modern medicine. Moreover, that different disease-related goals of medicine dictate classificatory choices in a pragmatic way is part of the reasoning behind the ICD-10:

The *purpose* of the ICD is to permit systematic recording, analysis, interpretation and comparison of mortality and morbidity data collected in different countries or areas and at different times. The ICD is *used to translate diagnoses of diseases* and other health

problems from words into an alphanumeric code, which permits easy storage, retrieval and analysis of the data (World Health Organization 2016, 3, my emphasis).

What is more, a similar approach is seen in the pages of widely used medical textbooks, which provide different classificatory descriptions various diseases, and do so according to what aspect of a particular disease is being outlined (Kasper et al. 2015; Walker et al. 2014). What is noteworthy in proposing a broadly pragmatic approach to disease classification is that arguments about which approach is best, or which provides the most understanding for *all medicine* (Scadding 1959; Snider 2003), are rendered somewhat redundant.

What should be at the center of debates over disease classification is what kind of information about a disease is most useful for the goal at hand—what approach to classification one should adopt depends on their medical aims. To see this, consider a case where we have etiological, pathophysiological, and symptom-based classifications of a particular disease, say IHD. As mentioned above, the etiological classification of IHD provides the following risk factors: Smoking, physical inactivity, obesity, hypertension, and excessive alcohol consumption. A symptom-based classification of IHD includes as symptoms extreme fatigue, shortness of breath, lightheadedness or fainting, angina, heart palpitations, and edema, among others. And a pathophysiological classification of IHD states that it is as inadequate supply of blood and oxygen to a portion of the myocardium due to both the fixed vessel narrowing and abnormal vascular tone as a result of atherosclerosis and endothelial dysfunction.

Now say our goal is to outline strategies for the *primary prevention* of IHD. Such strategies may include programs to encourage exercise, a healthy diet, and to discourage smoking. While we have classifications of the disease according to its observable symptoms and its underlying pathophysiology, such information is not useful for primary preventive strategies. If we are concerned with preventing the onset of the disease, we should focus on its distal causes. These are provided by the disease's etiological classification. Likewise, if our goal is to formulate a *diagnostic hypothesis* for a given condition, we can focus on observable symptoms. If one were to come in for a consultation complaining about fatigue, chest pain, occasional dizziness, and swelling in their legs, a doctor could hypothesize that they suffer from IHD. On the other hand, as mentioned in Section 1.3 above, *diagnostic confirmation* may require symptomatic information that reveals pathophysiological facts about the condition, and thus the pathophysiological approach is most appropriate. In each case, we have access to different classificatory approaches for IHD, however, certain approaches are more applicable than others given the goal at hand.

Another key concern in this regard centers on which approach is best for predicting the effectiveness of medical interventions. Medical interventions are targeted at distal or proximal causes of disease. Given that interventions target these causes, researchers may want to utilize etiological or pathophysiological classifications for therapeutic prediction. However, as I outline in Section 1.4, the pathophysiological approach can include information about the distal causes of a disease while

providing information about more proximal causes. This renders the etiological approach relatively redundant for therapeutic prediction.

The main reason for preferring the pathophysiological approach in the context of predicting medical effectiveness, however, is related to the information it provides about mechanisms. There are numerous arguments in the philosophy of medicine about the importance of mechanisms, mechanistic reasoning, and mechanistic evidence for causal inference in medicine (Russo and Williamson 2007; Illari 2011; Clarke et al. 2014; Parkkinen et al. 2018 (see Chapter 2 below)). Briefly, it is argued that mechanistic information is often necessary for such inferences in medicine because (a) it tells us how a cause brings about an effect, and (b) it can help us to avoid accepting problematic statistical results.³ This line of reasoning has been extended to therapeutic prediction (Steel 2008; Cartwright 2012; Fuller and Flores 2015). Since predictions of medical effectiveness are causal inferences about interventions that will be used in practical contexts, we need to be able to generalize from the restricted scenarios in which statistical evidence is commonly produced—the experimental context—to target contexts where such interventions will be administered. Interventions can fail in target contexts for many reasons, one of which is differences between relevant mechanistic features of the experimental context and target context. Thus, to generalize more reliably we should have mechanistic evidence about how interventions may work, and how they may fail to work, in target contexts. One aspect of this is having information about the mechanisms of the disease our intervention is targeting. Since pathophysiological classifications provide mechanistic information, it stands to reason that we should expect the pathophysiological approach to be helpful for predicting medical effectiveness.

1.6 Conclusion

The first goal of this Chapter was to describe and clarify three prominent approaches to disease classification: The etiological approach, the symptom-based approach, and the pathophysiological approach. The etiological approach classifies a disease according to its cause or causes. There are two models that fall under the etiological approach. The first, the monocausal model, is well-suited for classifying diseases that are thought to have one necessary and universal cause, such as infectious diseases. The second, the multifactorial model, is more appropriate for classifying diseases which can be caused by multiple constellations of causes, such as most chronic diseases. While these two models may not be as distinct as previously thought, they each have benefits for treatment development and the understanding of risk factors, respectively.

The second approach I examine in this Chapter is the symptom-based approach. This approach classifies diseases according to their symptoms or clinical pictures. I argue that symptom-based classifications can be problematic when used for making decisions about treatment strategies, unless the symptoms cited in the classification can help reveal relevant information about the underlying

³ For concrete examples of (b), see Biddle (2007), Jukola (2017), and Goldacre (2012).

pathophysiology of the disease in question. However, the symptom-based approach is useful in some contexts, particularly in the realm of diagnosis.

The final approach I discuss is the pathophysiological approach. This approach classifies diseases according to their constitutive basis. I argue that this constitutive basis is typically articulated in mechanistic terms. We are concerned with the dysfunction of the parts and processes in the body when we speak of disease. There are two forms the pathophysiological approach may take: Classification by structure, or classification by function. I argue that the latter, the more pathophysiological approach can include information about the former without sacrificing its focus on function. This, I argue, is a strength of the pathophysiological approach.

I then argued that medicine should adopt a broadly pragmatic approach to disease classification. The medical goal in question should dictate the classificatory approach used. This is because each approach had its own advantages and disadvantages that can affect our capacity to achieve certain goals. I close by describing how therapeutic prediction would benefit most from information provided by pathophysiological classification of diseases.

Chapter 2

Integrating Mechanistic Evidence for Inferring Medical Effectiveness

2.1 Introduction

Two competing views have emerged in debates about the relationship between evidence and causal inference in medical research. The first, the *statistical approach*, argues that good statistical evidence from randomized control trials (RCTs) or meta-analyses, is necessary and sufficient to warrant causal claims. The second, the *mechanistic approach*, suggests that both statistical and mechanistic evidence are necessary. This latter argument is often made on the grounds that *prediction* and *extrapolation* require knowing the underlying mechanisms of a purported causal relationship. The view that mechanistic evidence is important for extrapolating causal claims from one context to another is well-established, however, there is less in terms of how such evidence can or should be integrated with statistical evidence in service of this goal. In this Chapter, I clarify the concept of mechanistic evidence; argue that a key distinction between the mechanistic and statistical approaches lies in assumptions regarding the probability of there being differences between contexts; and offer an informal model for integrating available mechanistic evidence with statistical evidence for predicting medical effectiveness.

In the context of medicine, the debate often features in discussions about what evidence is required to warrant causal claims in medical research. The statistical approach argues that all that is required is evidence from RCTs or meta-analyses (EBM Working Group 1992; Howick 2011a). Its most prominent support comes from the evidence-based medicine (EBM) movement, as evident in several of the discipline's evidence hierarchies, which typically consider evidence gathered from RCTs and meta-analyses as the highest quality available (Guyatt et al. 2008; OCEBM Levels of Evidence Working Group 2011). Additionally, some regulatory bodies, such as the US Food and Drug Administration (FDA) (see Stegenga 2017), adopt this approach in their approval processes for new pharmaceuticals. A key feature of this view is that if there is good evidence from RCTs or meta-analyses, that is statistical evidence, then evidence from other sources, such as mechanistic evidence, is expendable.

Recently, many philosophers and medical practitioners have challenged the central doctrines of the EBM movement. Broadly speaking, some have questioned the rationale behind evidence hierarchies (Upshur 2005; Solomon 2011; Stegenga 2014). More specifically, some have challenged the status of RCTs as the benchmark of evidence quality (Worrall 2002; Doherty 2005; Cartwright 2007) while others have challenged the notion that meta-analyses produce objective amalgamated

results (e.g. Stegenga 2011, Jukola 2015). A common thread through many of these critiques is an emphasis on the shortcomings of statistical evidence.

Given these concerns, some have argued that statistical evidence is insufficient to warrant causal claims in medicine and have suggested that it should be augmented with mechanistic evidence (Russo and Williamson 2007; Clarke et al. 2014). Among this mechanistic approach's most ardent defenders are proponents of EBM+ movement, who argue that, provided it meets certain standards, mechanistic evidence should be considered equal in quality to statistical evidence (Parkkinen et al. 2018). This, some claim, is especially the case given that mechanistic evidence mitigates certain potential intrinsic biases of statistical evidence, and vice versa (Illari 2011). Additionally, some professional organizations, at least in principle, promote this view. The International Agency for Research on Cancer, for instance, outlines the importance of mechanistic evidence in its guidelines for the identification of carcinogenic hazards (IARC 2019), while the National Institute for Health Research encourages studies investigating mechanisms (NIHR 2019).

Understanding the differences between these two approaches to evidence is key to explaining the relation between evidence and causal inference, particularly in the context of predicting medical effectiveness. Highlighting these important differences sheds light on the underlying reasons behind excluding or including mechanistic evidence from such predictions. Here, I argue that the difference between the approaches is not just the attitude each adopts toward the quality of mechanistic evidence. I show that, in fact, several proponents of the statistical approach see mechanistic evidence as useful in some cases. Another crucial difference, when it comes to predicting effectiveness, is in disparate assumptions regarding the probability that the context to which a drug is being applied is different in relevant respects to the context in which the drug was tested. I argue that the standard view of the mechanistic approach is that this probability is high, and that mechanistic evidence can help in assessing the extent of these differences.

However, this raises the important question of how available mechanistic information should be integrated with statistical evidence. Much has been said about the difficulties associated with amalgamating evidence (e.g., Douglas 2012; Stegenga 2013; Fletcher et al. 2019). Still, this question has recently received increased attention, with some proposing the use of Bayesian networks (Landes et al. 2018; Marchionni and Reijula 2019). Using Toulmin's (2003) model of argumentation as a base, I suggest a schema for the systematic integration of mechanistic evidence with statistical evidence for predicting medical effectiveness. This model, I argue, is not necessarily an alternative to the more formal Bayesian network approaches, but it does avoid several of the issues they face.

I begin in Section 2.2 by offering an account of *mechanistic evidence* to work with throughout the Chapter. Mechanistic evidence is often defined in contrast to statistical evidence. Here, I follow Marchionni and Reijula's (2019) claim that this strategy is largely problematic, particularly when making singular conceptual comparisons between the two types of evidence. However, I argue that taken together, these contrasts afford us insight into the nature of mechanistic evidence. Following this, in Section 2.2.2, I explain the relation between mechanistic reasoning and mechanistic evidence

before turning to my comparison of the statistical and mechanistic approaches to evidence and causal inference in Section 2.3. I argue that, in the context of predicting medical effectiveness, a crucial distinction lies in each approach's assumptions regarding contextual similarity (Section 2.3.2). I then briefly outline the kinds of differences between contexts that would be relevant to predictions of a drug's effectiveness. In Section 2.4, I present my informal model of evidence integration. I explain the key features of the Toulmin model of argumentation. Following this, I use the case of the failed deployment of the HIV anti-retroviral drug efavirenz in Zimbabwe to demonstrate how the model can be applied to integrate mechanistic evidence with statistical evidence. In Section 2.5, I address a potential objection to the use of the Toulmin model before concluding.

2.2 What is Mechanistic Evidence?

It is important to define mechanistic evidence before tackling the problem of how integrate it with statistical evidence. Given the nature of the debate, it is common to describe mechanistic evidence by contrasting it with statistical evidence. This contrast has been characterized in several ways, including along methodological, numerical, populational, abstractive, and ontological lines. Marchionni and Reijula (2019: 57-58) have provided detailed critiques for the first four of these, which I summarize below. I argue that the last on its own, a distinction along ontological lines, is just as untenable. Following this, I provide an account of mechanistic evidence, arguing that it exhibits *appropriate-level specific* information about the components of causal pathway in question. Another important issue involves the relation between mechanistic evidence and *mechanistic reasoning*. While these terms are often conflated, I show that they are different in important ways, and explain that mechanistic reasoning relies, in part, on having mechanistic evidence.

2.2.1 *Contrasting Statistical and Mechanistic Evidence*

Distinguishing between statistical and mechanistic evidence by appealing to differences in methodology (Russo and Williamson 2007; Gillies 2018) is questionable. Statistical methods regularly provide information about mechanisms, while purportedly mechanistic methods, such as wet or dry lab work, often makes use of and provide statistical data. Moreover, it is not clear that what sets mechanistic evidence apart from statistical evidence is that the former is necessarily qualitative whereas the latter is quantitative (Waldner 2012). We can, and regularly do, use quantitative data to make inferences about mechanisms, while qualitative features of the world are used to stratify groups for statistical research. Relatedly, arguing that statistical evidence provides information about populations, whereas mechanistic evidence is about individuals (*ibid.*) is untenable since mechanistic information can apply to groups just as much as it could to individuals. Finally, *strictly* characterizing statistical evidence as providing macro-level information and mechanistic evidence as providing micro-level information (Claveau 2012) is problematic simply because mechanisms can be described at various levels of abstraction and grain.

Another way in which the distinction between statistical evidence and mechanistic evidence has been made, and one worth considering in more depth, is as evidence of entirely different *things*—an *ontological distinction*. Illari (2011), in response to what she sees as ambiguity in the RWT, argues that statistical evidence is evidence of *correlation* whereas mechanistic evidence is, straightforwardly, evidence of *mechanisms*. That is to say, where former refers to evidence illustrating a difference-making relationship between two variables, the latter simply refers to “evidence of the entities or activities that make up mechanisms, or the organization of those entities and activities by which they produce the phenomenon the mechanism is known for” (Illari 2011: 145). This characterization has been carried over into the EBM+ movement.

This definition, however, is too permissive about what counts as mechanistic evidence. Since a mechanism is defined as entities and activities and their organization (Machamer et al. 2000), the definition would count as evidence of mechanism *anything* that substantiated the existence of such entities and activities organized in a productively continuous way.⁴ Say we postulate the existence of a particular mechanism M_1 connecting two variables C and E . Finding evidence of correlation V between the variables counts as (at least some) evidence for the existence of M_1 . Yet, if someone proposes that there is a different mechanism M_2 connecting C and E , then V counts as evidence for M_2 as well. Both M_1 and M_2 cite entities and activities and their organization, and since V gives evidence for both, it counts as evidence of mechanisms. So, it seems that (good) evidence of correlation is trivially evidence of mechanism, at least if we use the definition above.

Another reason evidence of correlation satisfies this definition is that this view is consistent with the claim that statistical evidence is evidence about *difference-making*, and mechanistic evidence is evidence of *productive continuity* between entities. However, if the different things we are getting evidence of are difference-making and productive continuity, then this view is not saying much at all about the difference between statistical and mechanistic evidence. In discussions of mechanisms, productive continuity is the notion that the relevant activities of one element give rise to the relevant activities of the next from the start to the completion of a mechanism (Machamer et al. 2000; Steel 2008). Difference-making, according to Woodward’s (2003) counterfactual-manipulationist account of causation, is the notion that ideal interventions on one variable C will lead to changes the value or probability-distribution of another variable E . The distinction between these two concepts is not so stark. If a mechanism exhibits productive continuity, then it is necessarily the case that changes in the relevant activities of particular element will lead to changes in the relevant activities of another. Similarly, if we know two variables exhibit a difference-making relation, then it is necessarily the case that there is a productively continuous mechanisms between the variables. Therefore, distinguishing between the two types of evidence along these lines alone is inadequate.

⁴ I adopt Machamer et al.’s (2000) definition, which is similar to Illari’s. I do not have the space to examine other accounts of mechanisms here (e.g. Bechtel and Abrahamson 2005; Gillies 2017). I direct readers to Anderson (2014a, 2014b) for a good overview of mechanisms in the philosophy of science.

2.2.2 *Mechanistic Evidence and Mechanistic Reasoning*

Overall, drawing a stark distinction between statistical and mechanistic evidence is difficult. There are numerous overlaps between the two types of evidence, which make defining mechanistic evidence in this way problematic. The contrasts outlined above are questionable when taken in isolation, however, when taken together, some provide useful insights into two important features of mechanistic evidence.

The first is that mechanistic evidence gives us information about particular components of a causal process between two variables. To see this requires recognizing that there *are* methodological considerations for obtaining mechanistic evidence. It is right that there is no broad methodological distinction between these two categories of evidence such that certain *types* of methods, say RCTs, strictly produce statistical evidence, while others, say wet lab work, strictly produce mechanistic evidence. However, we regularly tailor the methods we use to acquire information about one or the other type of evidence. In other words, *token* cases of research methodologies, such as an RCT tailored to investigate the influence of a particular gene mutation on drug response, can produce mechanistic evidence, and there are significant methodological constraints in these cases.

Similarly, ontological considerations are useful too. I have argued above that the ontological distinction is problematic because it is too vague. However, this also helps us recognize that some correlational data provides information about the components of the mechanism between two variables. This, in conjunction with methodological considerations above, helps us recognize that mechanistic evidence is not simply evidence that *some* mechanism exists; statistical evidence, evidence of correlation, whatever we call it, already provides that for us. Rather, mechanistic evidence exhibits *specificity*—which is to say, it provides information about *specific* entities, *specific* activities, and/or *specific* organizational structures of the causal process between two variables of interest.

Understanding that mechanistic evidence exhibits specificity helps us appreciate the second feature of mechanistic evidence: It exhibits specificity at an *appropriate granular and populational level*. The problems with the macro-level-micro-level and population-individual contrasts of statistical and mechanistic evidence together demonstrate that there is no fixed level of specificity for mechanistic evidence. We might, for instance, want mechanistic evidence regarding genetic mutations for an individual, or we might want mechanistic evidence about the genetic make-up of a target population. Indeed, the specificity of mechanistic evidence varies according to the level of abstraction we choose (Cartwright 2012). However, the level of specificity is *appropriate* when it serves the goal in question. While information about the genetic make-up of a target population may be appropriate for predicting the effects of the rollout of a particular drug, it serves little purpose for estimating the drug response of a given individual. Thus, mechanistic evidence is appropriate-level information about the specific entities, specific activities, and specific organizational structures of the causal pathway between two variables of interest.

Of course, appropriate-level specific information about components of a causal pathway certainly is not all always necessary for making predictions about medical effectiveness. However, as I argue in Section 2.3, assumptions about the probability of there being differences in relevant causal features between a sample and target context play a role in the arguments for excluding or including available mechanistic evidence in inferences of medical effectiveness. One immediate implication here is that the propriety of available mechanistic evidence should be evaluated before its inclusion in such inferences (Parkkinen et al. 2018).

Another important issue facing the inclusion of mechanistic evidence is its relationship with mechanistic reasoning. Mechanistic reasoning has been characterized as inferences from mechanistic knowledge to claims that an intervention has a particular effect, with explicit linking of the intervention to its outcome (Howick et al. 2010; Marchionni and Reijula 2019). However, given the numerous areas of medical research which feature reasoning about mechanisms, this is too narrow a definition. A vast number of inferences in medicine involve reasoning about mechanisms including hypothesizing the existence of specific mechanisms, inferring how a particular intervention produces an outcome, and developing drugs based on their mechanistic features. Simply put, mechanistic reasoning is not limited to inferences about the effects of interventions.

Mechanistic reasoning, then, is inferences about the underlying mechanism of a given phenomenon. It is important to note that the product of these inferences is not mechanistic *knowledge*, but rather mechanistic *theories* or mechanistic *claims* (Campaner 2011). For example, the hypothesized disease mechanisms of *C. auris* are inferred using well-established theories and evidence about closely related pathogenic species within the *Candida* clade, such as *C. albicans* and *C. glabrata* (Chatterjee et al. 2015). Such reasoning may be based on existing hypotheses regarding underlying causal structures, theories about mechanisms, or, notably, mechanistic evidence. In the case of *C. auris*, researchers are concerned with looking for genomic information which confirms the hypothesized disease mechanisms of the fungus (*ibid.*). This illustrates how (good) mechanistic reasoning is reliant on (good) mechanistic evidence.

To reiterate, mechanistic evidence provides appropriate-level information about the specific entities, specific activities, and specific organizational structures of the causal pathway between two variables of interest. Relatedly, mechanistic reasoning involves making inferences about the underlying mechanisms of a particular phenomenon. Having characterized both mechanistic evidence and mechanistic reasoning, I now turn to comparing the statistical and mechanistic approaches to evidence.

2.3 Medical Evidence, Contextual Similarity, and Causal Consistency

This section focuses on comparing the statistical and mechanistic approaches to evidence. I begin by explaining their distinct attitudes towards mechanistic evidence before situating them in the context of predicting medical effectiveness. In that context, I argue that a key factor explaining the attitudes

toward evidence is each approach's assumptions regarding contextual similarity. The section ends with my account of *casual consistency*, which aims at articulating the kinds of differences between contexts that are relevant when making predictions about drug effectiveness.

2.3.1 *Two Attitudes Toward Medical Evidence*

Proponents of the statistical approach claim that evidence obtained from RCTs and meta-analyses is enough to establish the efficacy of a medical intervention (Howick 2011a). Generally, the EBM movement defends this position. This is illustrated by the EBM Working Group (1992), which claims that such clinical research should be favored over clinical experience and mechanistic reasoning. The statistical approach to medical evidence has been further ratified in several evidence hierarchies within the EBM, which typically rank evidence gathered from RCTs and meta-analyses above other evidence, such as that gathered through non-randomized trials, case reports, and importantly, mechanism-based reasoning (Guyatt et al. 2008; OCEBM Levels of Evidence Working Group 2011). Because of the view that statistical evidence, on its own, is necessary and *sufficient* to warrant causal inferences, I refer to proponents of this approach as *evidential monists*.

This cynicism towards mechanistic reasoning is typically defended on the grounds that it often results in erroneous causal inferences. A common argument here is that, because of their complexity, we often cannot have adequate knowledge of physiological mechanisms for mechanistic reasoning to be reliable (Howick 2011a: 136-143; Anderson 2012). Stegenga (n.d.) calls this the *trouble-in-the-box* argument for the statistical approach. This problem is often illustrated using the example of the CAST trial designed to test the efficacy of antiarrhythmic drugs (Greene 1992). It was (mechanistically) reasoned that because some people died due to cardiac arrhythmias within a year of myocardial infarction, antiarrhythmic drugs would help reduce one's risk of dying after suffering a heart attack. The CAST trial, however, demonstrated that such drugs increase overall mortality. This, according to proponents of the statistical approach, shows how mechanistic reasoning can lead to flawed and, at times, devastating results.

It should be noted, however, that the statistical approach does not completely disvalue mechanistic reasoning. Howick (2011a, 2011b), for example, states that it can be useful to make inferences from well-understood, simple mechanisms. Moreover, it has been suggested that, in the absence of adequate evidence, mechanistic reasoning is necessary for solving clinical problems (EBM Working Group 1992). Nevertheless, these considerations are consistent with proponents of the statistical approach's claim that mechanistic reasoning is neither necessary, nor sufficient for warranting casual claims.

The skepticism towards mechanistic reasoning has led to a depreciation of mechanistic evidence in the EBM movement. As mentioned in the previous section, reasoning from mechanisms depends on evidence for those mechanisms. Because of this, research which produces such evidence is relegated to the lower tiers of evidence hierarchies. Accordingly, a striking feature of the statistical approach, is the claim that if there is good evidence from RCTs or meta-analyses, then mechanistic evidence can be wholly disregarded when making claims about an intervention's effectiveness.

In opposition to this, proponents of the mechanistic approach argue that statistical evidence is necessary but insufficient to warrant causal claims in medicine and have suggested that available mechanistic evidence should always be included for such inferences (Illari 2011; Clarke et al. 2014). Currently, the most prominent defenders of this approach are those in the EBM+ movement, who “evaluating evidence of mechanisms *alongside* evidence of correlation” (Parkkinen et al. 2018: 20). These arguments can broadly be understood as defenses of the Russo-Williamson Thesis (RWT) which states that scientists require *both correlative and mechanistic* support for making causal claims (Russo and Williamson 2007). Given this strict adherence to the RWT, I refer to this stricter view of medical evidence as *evidential dualism*—the view that we *must have both* statistical and mechanistic evidence to warrant a causal claim. However, not all proponents of the mechanistic approach claim that the RWT is true in every case; I refer to this more amenable approach to the medical evidence as *evidential pluralism*—the view that sometimes one or the other kind of evidence is enough to warrant a causal claim. However, both evidential dualists and evidential pluralists argue for the importance of taking available mechanistic evidence into account when inferring causal claims in medicine. That is, provided it meets certain standards, mechanistic evidence should be considered equal to evidence gathered from RCTs and meta-analyses in evidence hierarchies and used to complement the evidence gathered in medical trials.

Proponents of the EBM+ movement have offered plausible grounds for the claim that the statistical and mechanistic evidence complement one another. For instance, some claim that mechanistic evidence mitigates certain potential intrinsic biases of statistical evidence, and vice versa (Illari 2011). Relatedly, others have highlighted the numerous historical cases where neither statistical evidence nor mechanistic evidence on their own were sufficient for accepting causal claims (Clark et al. 2014). The mechanistic approach is therefore consistent with the view that incorporating different types of accurate evidence will improve the inferences we make about the causal relation in question. In other words, if good mechanistic evidence is available, then including it will improve predictions about medical effectiveness.

The above discussion illustrates a broad contrast between these two approaches to evidence and causal inference. Because evidential monists regard mechanistic evidence as lower in quality in comparison to statistical evidence, the statistical approach adopts a *dismissive attitude* toward evidence. This is the notion that if evidence from RCTs or meta-analyses is available, then mechanistic evidence, can be disregarded when inferring that an intervention will be effective in some target context. In contrast, because both evidential dualists and evidential pluralists see statistical and mechanistic evidence as being equal in terms of evidential quality, the broader mechanistic approach embraces an *augmentive attitude* toward evidence. This is the notion that statistical evidence and mechanistic evidence complement one another, thus strengthening causal inferences. This entails that if mechanistic evidence is available, then it should be combined with statistical evidence when making causal claims.

While these opposing attitudes and considerations of evidence quality help explain the contrast between the statistical and mechanistic approaches in general, it is important to evaluate the further reasons for these different views in the narrower context of making predictions about the effectiveness of medical interventions.

2.3.2 Assumptions Regarding Contextual Similarity

When it comes to extrapolating the results of RCTs and meta-analyses to target contexts, advocates on both sides of the debate agree that if a sample and target are not sufficiently similar in relevant respects, then the results from the sample will not obtain in the target. The view that mechanistic evidence is inferior to statistical evidence only partially explains the dismissive view in the context of prediction. Here, I show that another reason, cited by proponents of the statistical approach, is that there is a low probability that a sample and target are different in relevant respects. The mechanistic approach, on the other hand, assumes a high probability that the relevant features of two contexts are different. This contrast has not been scrutinized to the same extent as the disagreements about evidence quality, but it is important in understanding why mechanistic evidence is excluded by evidential monists when predicting medical effectiveness.

Proponents of the statistical approach claim that results from RCTs can be extrapolated to a target context unless there is some reason to believe that the sample and target differ. Prominent EBM epidemiologists recommend that we should deal “with the issue of generalizability by accepting that randomized trials apply to wide contexts unless there is compelling reason to believe the results would differ substantially as a function of particular characteristics of those patients” (Post et al. 2013: 641-642). Steel (2008) refers to this as *simple extrapolation*, however, a more accurate characterization is provided by Stegenga (2015), who refers to as *simple extrapolation unless* (SEU) and notes that this view is shared by other EBM practitioners and researchers (see Guyatt et al. 2015; Moher et al. 2010). Moreover, SEU is actively taught to future EBM researchers and clinicians (see Straus et al. 2019: 126-127).

Importantly, the statistical approach argues that reasons to doubt that the results from a sample context can be applied to a target will *rarely be found*. Guyatt et al. (2015: 116), for example, claim that “you will usually not find a compelling reason, in which case you can generalize the results to your patient with confidence”. There are at least two possible motives for this rationale among EBM practitioners.

First, they may believe that even if we conduct mechanistic studies (and thus use mechanistic reasoning) and happen to find evidence that seems to cast doubt on our extrapolation, that evidence would fail to constitute a “compelling reason”. This is because, according to the proponents of EBM, mechanistic reasoning (and thus mechanistic studies) is a low-quality source of evidence. Strikingly however, many of the same authors, including Post et al. (2013), argue that compelling reasons may be found by evaluating pathophysiological, socioeconomic, and epidemiological characteristics. A recent textbook by leaders in the movement states that we should “consider whether our patient’s socioeconomic features or pathobiology are so different from those in the study that its results are

useless to us and our patient; only then should we discard its results and resume our search for relevant evidence” (Straus et al. 2019: 126). Given that mechanistic studies are often used to find such evidence, the mistrust of mechanistic evidence and reasoning may play less of a role than some may think.

The second possible motive behind the claim that compelling reasons are rarely found, is the belief that relevant differences are simply unlikely. That is, proponents of SEU may believe that the probability of there being relevant differences between a sample context and the target context to which the results are being applied is low enough that we can typically extrapolate results. There is evidence that this is ultimately the rationale behind the statistical approach. Post et al. (2013), for example, consider adverse effects of treatments to be unlikely and Straus et al. (2019: 127) claim that “differences in response are extremely rare”. More explicitly, the influential GRADE Working Group writes:

In general, one should not rate down for population differences unless one has compelling reason to think that the biology in the population of interest is so different from that of the population tested that the magnitude of effect will differ substantially. Most often, this will not be the case. (Guyatt et al. 2011: 1304)

This demonstrates that proponents of the statistical approach adopt an *equative view* of contexts and their populations—there is a low probability that a sample and target context differ in relevant respects. This contributes to dismissing mechanistic evidence when it comes to predicting medical effectiveness. Given the high probability of relevant similarity between a sample and target, mechanistic evidence can be dismissed when applying the results of an RCT to the target.

Proponents of SEU provide evidence in support of the equative view. Post et al. (2013: 641) cite several studies showing that relative effect measures are generally consistent across RCTs included in meta-analyses regardless of differences in baseline risk of illness and factors such as participant age or sex.⁵ At the same time, they acknowledge that there are cases where such factors do impact effect sizes.⁶ and thus, we should look “carefully for evidence (preferably from RCTs) that suggests differential effects for the subgroup of interest and applying appropriate criteria in judging the likelihood that a subgroup effect is real” (*ibid.*). Fuller (2013), however, brings up a striking issue regarding this evidence: It only demonstrates consistency in relative effect measures across RCTs and not the generalizability, or external validity, of those results. In other words, it shows that the equative view is valid when it comes to the populations and contexts of RCTS. The problem here, as proponents of the mechanistic approach argue (see below), is that RCTs typically enforce strict criteria on participant inclusion and exclusion and thus they run the risk of being

⁵ The references Post et al. (2013) provide are: Schmid et al. (1998), Furukawa et al. (2002), Deeks (2002), Turnbull, Neal et al. (2008), and Turnbull, Woodward et al. (2008).

⁶ Here, Post et al. (2013) cite Hlatky et al. (2009) and Pignon et al. (2009).

unrepresentative of populations outside of those studies. There is always a chance that populations not constrained by the selection criteria of RCTs will differ in relevant ways.

The above studies are not the only evidence provided by Post et al. (2013) for the equative view. They claim that a Cochrane review (Odgaard-Jensen et al. 2011) concluded that “participation in RCTs is associated with similar outcomes to receiving the same treatment outside RCTs” (Post et al. 2013: 641). They suggest that this goes against the notion that results from RCTs are not generalizable. However, a careful reading of the Cochrane review in question reveals that no such conclusion is drawn. In fact, the review compares control trials with randomized allocation of participants to control trials without randomization and concludes that “the results of randomised and non-randomised studies sometimes differed. In some instances non-randomised studies yielded larger estimates of effect and in other instances randomised trials yielded larger estimates of effect” (Odgaard-Jensen et al. 2011: 7). It is unclear how these conclusions entail that we should expect that the outcomes in non-RCT settings will be similar to those observed in RCTs.

Nevertheless, a more recent Cochrane review may count towards the equative view. Anglemyer et al. (2014) compared the measured effects of interventions in RCTs with those of well-conducted observational studies and found that there was little difference between them. It may be concluded from this that the results from RCTs are at least generalizable to the less constrained settings of observational studies, and thus there is some warrant for the equative view. Yet, the authors are clear that there can be differences between the effect sizes measured in RCTs and those measured in observational studies, it is just that those differences are not likely the result of differences in study design (Anglemyer et al. 2014: 15). Indeed, this fits with the proposal of those in the EBM movement that good reasons for doubting the generalizability of RCTs can be found by evaluating pathophysiological, socioeconomic, and epidemiological characteristics of the populations in questions. These are what constitute relevant differences between sample and target contexts.

Evidential dualists and pluralists, too, argue that relevant differences between a sample and target entails that the results from the sample will not hold in the target (see Parkkinen et al. 2018; Cartwright 2012). Moreover, they agree that pathophysiological, socioeconomic, and epidemiological differences could lead to different treatment responses. For instance, those in EBM+ movement cast the issue as one of external validity and extrapolation and argue that behavioral and biological differences between a sample and a target context may lead to harmful or diminished responses to treatments in the target population (see Clarke et al. 2014; Parkkinen et al. 2018).

Where the mechanistic approach diverges from the statistical approach is in its assumptions about the probability of there being differences between a sample and target context. They assume that there is a high probability of there being differences between contexts. This, it is argued, is a result of typically strict inclusion and exclusion criteria set for clinical trials. Parkkinen et al. (2018: 5) claim that “a study population is typically highly idealized, and thus differs from the target population in important ways”. And in his explanation of SEU, Stegenga states:

Given the large number of criteria that many clinical trials employ which stipulate the properties that a potential subject must have (and other criteria which they cannot have) to be included in the trial, there are almost always differences between a particular real-world patient and the subjects in a clinical trial (2015: 69).

Proponents of the mechanistic approach thus adopt a *differential view* of contexts and their populations—there is a high probability that a sample and target context differ in relevant respects.

This differential view is a largely overlooked motive behind the evidential dualist's claim that mechanistic evidence should be considered alongside statistical evidence. It is not just that both types of evidence augment one another, but also because some characteristics of clinical trial methodology (e.g., participant inclusion and exclusion criteria) make it very likely that the features of target contexts are different in some relevant way to those of the sample. If relevant differences mean that the intervention will not have the same effect in the target as it did in the sample, and the probability of relevant differences between a sample and target are high, then, so the argument goes, having evidence about those relevant differences will improve predictions about the effects of treatments. For the evidential pluralist, this information comes in the form of mechanistic evidence:

it is almost never the case that clinical studies in the study population will directly establish both a suitable association and mechanism that will apply to the target population ... one typically needs to consider evidence of mechanisms arising from sources other than the clinical studies that establish a correlation in the study population. (Parkkinen et al. 2018: 25)

Given the claim that we need evidence about the relevant differences between a sample and target to improve our predictions, it would help to know what kinds of features may be relevant across contexts.

2.3.3 Causal Consistency

The essential concern in the debate about evidence in the context of medical effectiveness is whether results from randomized studies and meta-analyses can be reliably applied to target contexts. For proponents of the mechanistic approach, the difficulty is that once a causal relationship between an intervention I and an outcome O (I - O relation) is well-established in a sample setting \mathbf{S} (say by a well-designed RCT), all we have is evidence of a causal law, or causal model, governing the relation *in that setting*. Put another way, we have evidence of *efficacy*—the measure of the causal strength of an intervention in a controlled environment (see the Introduction of this dissertation). What limits our confidence to the sample context is (i) its reliance on relevant contextual features of \mathbf{S} and, importantly, (ii) that \mathbf{S} and a target setting \mathbf{T} are *likely differ with respect to these features*.

For evidential dualists and pluralists, accepting (i) and (ii) above entails that we require evidence of what the features on which the I - O relation relies in \mathbf{S} are, and second, we need evidence that those features are present in \mathbf{T} . Without such evidence, our confidence in the claim that I will bring

about O in \mathbf{T} should be lower than our confidence that I will bring about O in \mathbf{S} . Put another way, in order to consider a prediction of medical effectiveness reliable, we need compelling reasons to think that the *relevant features* of the sample and target contexts are *sufficiently similar* (Parkkinen et al. 2018).⁷ What it means for two contexts to be sufficiently similar in relevant respects is a two-part question. First, we need to ask what the relevant features, which must be shared between the sample and target contexts, are. And second, we need an idea of what counts as “sufficiently similar” enough for a prediction to be reliable. In this section, I consider the first of these issues and how mechanistic evidence helps evaluate these features. The relevant features have been well-theorized and explicated by others, most notably Cartwright (2012), so I will provide only a brief overview of my own. My motivation for putting these features into my own terms here, besides what I see as being a simpler characterization, is so they can more easily be included in the heuristic developed in the next section.

When it comes to predicting the effectiveness of medical interventions, what we are concerned with is whether a causal law or causal model, which we have established governs a relationship between two variables in one context, governs the same relationship in another. This idea can be captured by the idea of *causal consistency*. That is, the prediction that I , which brings about O in \mathbf{S} , will be effective at bringing about O in \mathbf{T} , is the claim that \mathbf{S} and \mathbf{T} are causally consistent with respect to the I – O relation. Two contexts are causally consistent with respect to a given causal relation to the extent that they share causally relevant properties. There are three properties of causal consistency:

Causal structure: The causal structure of the I – O relation in the \mathbf{S} and the causal structure of the I – O relation in \mathbf{T} are consistent with one another.

Causal enablers: The factors which enable the I – O relation in \mathbf{S} are present in \mathbf{T} .

Causal inhibitors: Factors which inhibit the I – O relation, which are absent in \mathbf{S} , are also absent in \mathbf{T} .

Evaluating these properties is key to assessing overall causal consistency across two contexts. Evaluating the first, causal structure, involves directly comparing the components of the causal pathway from I to O in the sample context to the components of the causal pathway from I to O in the target context. The second and third are features of the context which may influence the components of the causal pathway such that the I – O relation is modified in some way. Causal enablers are based on the reasoning that the I – O relation does not depend solely on the introduction of I but also on other factors, external to its causal structure, without which, I would not cause O .⁸

⁷ In Chapter 4, I refer to this approach as *extrapolation by sufficient similarity*.

⁸ This point is well known in philosophical discussions about interventions in social sciences and medicine. Cartwright makes the point by appeal to Mackie’s (1965) INUS causes to clarify what she calls having “the right support team” (2012: 976); Stegenga expresses it as part of ignoring background knowledge about how an intervention works when

Evaluating this property involves making sure that the relevant background conditions of a target context are sufficiently similar to those which enable the intervention to bring about its outcome in the sample context. Assessing causal inhibitors requires having information about those factors which could obstruct the $I-O$ relation. If these factors, which were not present in \mathbf{S} are present in \mathbf{T} , then there is less chance that the $I-O$ relation will obtain.

The more homogenous these properties are across contexts, the more causally consistent the contexts. *Total causal consistency* refers to the case where all three properties are homogenous across separate contexts. This would ensure that an intervention which is effective in one context would be effective in the other. As noted, though, such consistency rarely, if ever, occurs. Yet, when extrapolating claims of effectiveness, it is often assumed that the two contexts in question are sufficiently causally consistent. Standard RCTs may be a good starting point for establishing such claims in target contexts but we need to evaluate causal consistency as well; if we have an idea of the extent to which two contexts are causally consistent, then we can revise our confidence in the suitability of the causal law or causal model for the target context and make better predictions.

Evaluating causal consistency involves finding evidence. We need evidence which increases (or decreases) our confidence in the homogeneity of causal structure, causal enablers, and causal inhibitors. Evidence of causal consistency comes in numerous forms and from any methodology associated with the different levels on evidence hierarchies. But mechanistic evidence, when available, is particularly well-suited for this task because it gives us the kind of information needed for evaluating each property of causal consistency. By providing appropriate-level information specific features of a causal process, mechanistic evidence gives us information about differences between the sample and context which will affect I 's capacity for producing O .

When it comes to *causal structure*, we are concerned with the components that make up the causal pathway from I to O . Plainly, mechanistic evidence is most suited for this. If we can establish that the mechanism of action of a drug in the sample context \mathbf{M}_S and the mechanism of action of the drug in the target context \mathbf{M}_T are consistent with respect to the $I-O$ relation, then we have good reason to believe that the underlying causal structures of both are consistent with one another. This means first gathering evidence for \mathbf{M}_S . This evidence should, as stated in Section 2.2, be about appropriate-level specific components of the causal pathway from I to O . Once we have evidence of \mathbf{M}_S we can gather evidence that the \mathbf{M}_T will be similar if the intervention is deployed in that context. If it is similar enough, then we can have increased confidence in our prediction. If there is evidence to the contrary, then we should decrease our confidence in the prediction.

For example, the CYP2D6 enzyme plays an important role in the mechanism of action of the pain reliever codeine. This gene is responsible for metabolizing the drug into morphine, its active metabolite which is required for its analgesic effect. However, people who carry two inactive copies of the CYP2B6 gene (various combinations of *4, *5, and *6 alleles) are far less likely to experience

extrapolating (2015: 69); and Marchionni and Reijula (2019) cover it in their robustness condition for the function of mechanistic evidence.

an adequate analgesic effect. This is because two inactive CYP2B6 genes are poor metabolizers of codeine and thus produce reduced morphine levels (Bhandari et al. 2011). Having this mechanistic information is crucial to predicting whether codeine will be an effective pain reliever for an individual patient. In cases like this, we are concerned with comparing the mechanism in the sample with the mechanism in the target to find differences in the actual mechanism.

Often, we either cannot gather, or have not yet collected sufficient evidence about all the components involved in M_S and M_T to do a complete comparison, which can be a problem for assessing *causal structure*. One proposed solution to these issues is to compare those stages of M_S and M_T where differences are most likely to occur (see Steel 2008). Another solution is to perform RCTs that have been specially designed to investigate the componential features of a proposed mechanism (Marchionni and Reijula 2019), which as I have outlined previously, produce mechanistic evidence.

At first it may seem strange to speak of *causal enablers* and *causal inhibitors* in the context of medical interventions which often involve biological mechanisms. These properties of causal consistency are seemingly more compatible with disciplines like economics and behavioral policy where background features of contexts have more obvious effects on the causal pathway between an intervention and an outcome. However, it is often the case that socioeconomic and epidemiological differences can indirectly affect drug response in target contexts. Furthermore, there are contextual pathobiological factors which are not directly associated with the mechanism itself, that may influence the components of the causal pathway between a drug and its intended outcome.

Appropriate-level mechanistic evidence, can help unpack how these features can influence the components of the causal pathways between I and O . Take β -lactam antibiotics for *S. aureus* infection. The mechanism of action for these drugs is well-established: β -lactam antibiotics inhibit synthesis of the peptidoglycan layer of the bacterial cell-wall. This happens by virtue of the antibiotics having a similar structure to d-alanyl-d-alanine, an amino acid ending on the peptidoglycan layer. The antibiotic targets and binds to active penicillin binding proteins which are essential for cell-wall synthesis. However, this mechanism of action has a better chance of occurring when the strain in question is methicillin-sensitive *S. aureus* (MSSA). Thus, if we find evidence that a person is infected with MSSA, then we can be quite confident that β -lactam antibiotics will be effective. If, on the other hand, we find evidence that the person is infected with methicillin-resistant *S. aureus* (MRSA), then we should have less confidence (but importantly, not zero confidence) that β -lactam antibiotics will be effective. This is because we have a well-established mechanistic theory that MRSA is resistant to β -lactam antibiotics. This resistance occurs due to the acquisition of a non-native *mecA* gene which encodes a penicillin-binding protein (PBP2a), with significantly lower affinity for β -lactams. This allows cell-wall biosynthesis to continue even in the presence a β -lactam antibiotic (Peacock and Paterson 2015).

Being infected with MRSA is a slightly different case of a biological factor influencing drug effectiveness to the one in the codeine example above. Here, instead of a component of the casual pathway itself being different, we are presented with a factor external to that pathway being

different, which inhibits the mechanism of action of the drug from occurring effectively. Mechanistic evidence helps us determine whether this causal inhibitor is absent or not and thus helps in predicting whether a *β-lactam* antibiotic will be effective for treating this case of *S. aureus* infection.

It is compelling that if appropriate mechanistic evidence is available, it should be included in our predictions about the medical effectiveness of an intervention for a target context. One issue facing the mechanistic approach, particularly those in the EBM+ movement is the lack of method for integrating mechanistic evidence with statistical evidence. If we are to accept the evidential pluralist's augmentive approach to evidence, then we ought to have methods for amalgamating these two types of evidence.

2.4 The Toulmin Model: A Schema for Integrating Mechanistic Evidence with Statistical Evidence.

In this section, I offer a schema for the systematic integration of mechanistic evidence with statistical evidence based on the Toulmin model of argumentation. The Toulmin model offers a simple, informal method for assessing causal consistency using available mechanistic evidence and evaluating the validity of predictions of medical effectiveness. I begin by explaining the key elements of the Toulmin Model and then outline how the model is applied to predictions of medical effectiveness. Following this, I use the case of the rollout of the HIV drug efavirenz in Zimbabwe to demonstrate how the model can be applied to integrate available mechanistic evidence with statistical evidence. This case had disastrous results that may have been avoided had relevant mechanistic evidence been included in policy considerations. I end the section by replying to the possible objection that Bayesian networks already provide a good method for amalgamating mechanistic evidence with statistical evidence.

2.4.1 *The Toulmin Model and Predicting Medical Effectiveness*

The Toulmin Model is a fixed procedural model, which, according to Toulmin (2003), all arguments must follow. The *simple Toulmin Model* consists of three elements (Figure 2.1). The first is a *claim* (C): the assertion, proposition, judgement, or view expressed in the argument. Since claims are open to challenges, those who make them should be in a position to provide a defense. Typically, this is done by producing or appealing facts or evidence on which the claim is based. The second element then, is *data* (D) (sometimes referred to as grounds) which support the claim being made. Naturally however, just how the data supports the claim can be questioned and so a third element is necessary: *Warrant* (W), which links the claim to the data. This can be a general principle such as “if D then C”, however it can also be articulated in more extended ways as required by substantial arguments. It might, for instance, be the case that D makes C more likely, or D entitles one to claim C. Consider the following simple example: Introducing more legislation to regulate motor-vehicle emissions will reduce overall carbon emissions in Canada (C). In Canada,

transportation is responsible for about 43 percent of overall greenhouse gas emissions (D). As part of the transport category, motor-vehicles contribute to Canada’s overall carbon emissions (W).

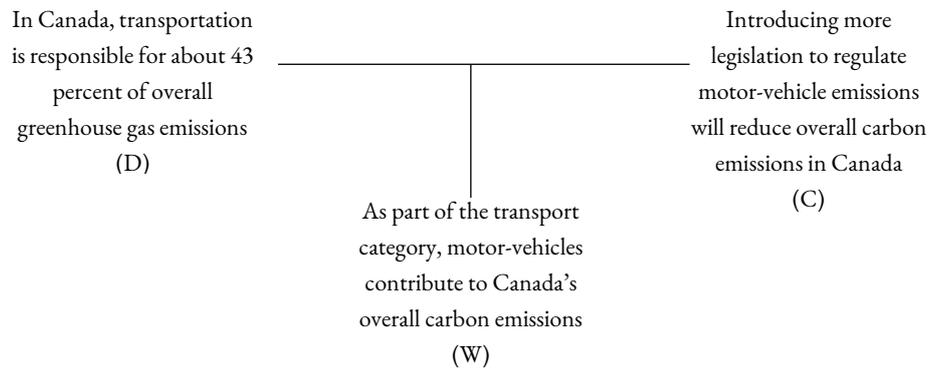


Figure 2.1. An example of the simple Toulmin Model

If the data is accurate and the warrant effectively links the data with the claim, then the claim is justified. However, if there were an exception to the claim, then a *rebuttal* (R) should be acknowledged and, if necessary, a condition of *qualification* (Q) included. Furthermore, if the warrant is questionable, then *backing* (B), which provides support for the warrant, is needed. These three elements, in addition to those of the simple Toulmin Model, constitute the *extended Toulmin Model*. Going back to the example above (see Figure 2.2), one may argue that vehicle sales may increase such that the effects of the legislation introduced are negated (R). Given this rebuttal, the claim needs to be qualified. The *force* of the qualification depends on the strength of the rebuttal. In this simple case, we can settle for a probabilistic qualifier. One may also question the warrant by arguing that the contribution to transport emissions from motor-vehicles is negligible and that other modes of transport, such as air travel and shipping, are the real problem. By providing the backing that motor-vehicles are responsible for approximately 85 percent of greenhouse gases in the transport category (B), the warrant is vindicated.

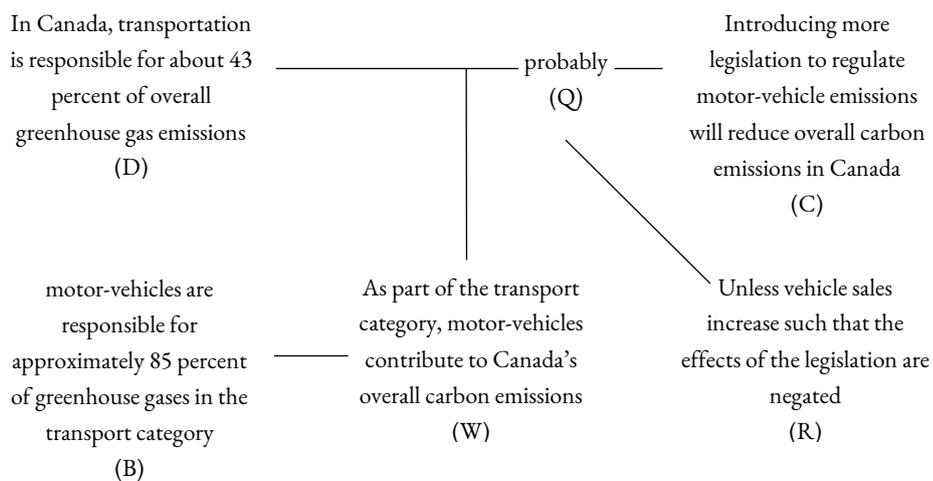


Figure 2.2. An example of the extended Toulmin Model

The elements of the simple Toulmin Model are present in every argument. The warrant, however, is often not overtly expressed, but rather implicit to the argument. Also note that further data can be used to support the warrant itself and thus serve as backing. Moreover, data can be used to support claims for or against a rebuttal to a given argument. The three elements of the extended Toulmin Model are also not always necessary. A backing is only required when the warrant is not immediately accepted. A qualifier is needed whenever there is a rebuttal to the argument. Conversely, sometimes in the absence of a rebuttal, a qualifier may still be necessary. With this, we have a general schema the Toulmin model (Figure 2.3).

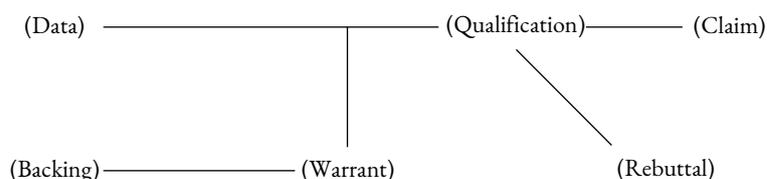


Figure 2.3. The general form of the Toulmin Model

The Toulmin model can be used to systematize available evidence to infer its overall support for a prediction of medical effectiveness. Consider the Toulmin model in Figure 2.4. Here, the claim is that the measure of effectiveness of a particular intervention I in target \mathbf{T} , Ef_T , will be the same as the efficacy measure established in a sample \mathbf{S} , Ef_S . The warrant for this claim is that \mathbf{S} and \mathbf{T} are sufficiently causally consistent which is supported by data in the form of evidence from well-designed RCTs or meta-analyses establishing Ef_S .

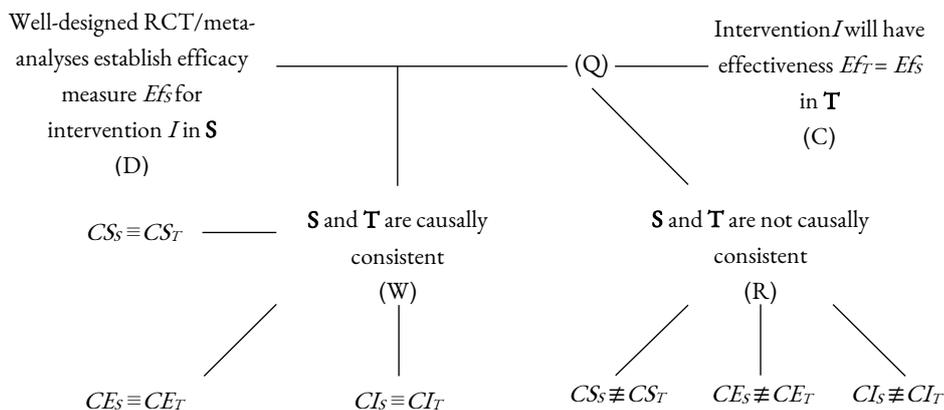


Figure 2.4. The Toulmin model for integrating mechanistic evidence with statistical evidence.

As mentioned above, the mechanistic approach's augmentive attitude toward evidence calls for the inclusion available mechanistic evidence. This is the backing for the warrant, and so should further support causal consistency by illustrating the homogeneity of its three properties: The causal structure in the sample is homogenous with the causal structure in the target ($CS_S \equiv CS_T$); the causal

enablers in the sample are present in the target ($CE_S \equiv CE_T$); and the causal inhibitors absent in the sample are also absent in the target ($CI_S \equiv CI_T$).

The mechanistic approach's dismissive view regarding contextual similarity captured in the qualifying rebuttal that **S** and **T** are not causally consistent. The qualification entailed by this rebuttal is determined by the inclusion of relevant mechanistic evidence supporting the heterogeneity of the three properties of causal consistency: The causal structure in the sample is heterogenous with the causal structure in the target ($CS_S \neq CS_T$); causal enablers in the sample are absent in the target ($CE_S \neq CE_T$); and causal inhibitors absent in the sample are present in the target ($CI_S \neq CI_T$).

Given that there is still ongoing debate between the statistical and mechanistic approaches (see Sections 2.1 and 2.3.1), one may justifiably ask *when* the Toulmin model should be employed. The evidential monist may hold that the model should rarely be used, while the evidential dualist may argue that it should always be used. Recall though, both evidential monists and dualists agree that if there were relevant differences between a sample context and target context, then that information would be found by looking at pathophysiological, socioeconomic, and epidemiological information about the case at hand. It seems reasonable then that even those in the EBM movement, who actively endorse the statistical approach and simple extrapolation unless, would accept that such a model could be used to explicitly demonstrate that there is no good reason to doubt the generalizability of a particular meta-analysis or RCT. Indeed, proponents of EBM admit that there are cases where the expected effectiveness of medical interventions fails to manifest in target populations. Even if these such instances are rare, the Toulmin model is helpful in demonstrating that a particular case is not one of those rare instances. The routine use of the Toulmin model then is not inconsistent with the central tenets of EBM.

To illustrate how the Toulmin model just described might be applied, I will use the case of the failed rollout of the HIV drug efavirenz in Zimbabwe. In the next subsection, I briefly outline the details of the case and, following this, briefly explain how the available mechanistic evidence could have been integrated with statistical evidence about the drug. Doing so may have helped convince policy makers to avoid deploying the treatment before establishing its safety, and thus may have mitigated the harms this caused.

2.4.2 Applying the Toulmin Model

There are several examples of cases where the consideration of mechanistic evidence affords us more accurate estimates of medical effectiveness. Some of these include variability in the effectiveness of codeine based on understanding how polymorphisms of the CYP2D6 genotype influence the metabolism of the drug (Madadi et al. 2013), the role of pulmonary, extrapulmonary, and behavioral factors in the treatment of severe asthma and chronic obstructive pulmonary disease (Agustí et al. 2017; McDonald et al. 2019), and dosage adjustment based on renal function and metabolism for the effective treatment of deep vein thrombosis with the drug enoxaparin (Nutescu et al. 2016). Another striking example, which clearly demonstrates the importance of considering available

mechanistic evidence has been thoroughly detailed by Park et al. (*forthcoming*): The case of efavirenz in Zimbabwe.⁹ Park et al. use this case to usefully demonstrate the importance of mechanistic reasoning and how Pearl and Bareinboim's (2014) do-calculus may be used to amalgamate statistical and mechanistic evidence. Given its applicability, and the thorough explanation provided by Park et al., I will use the Zimbabwe-efavirenz case to illustrate how the Toulmin model can be used to effectively organize existing evidence prior to such a quantitative analysis.

Efavirenz is an antiretroviral used in the treatment of HIV/AIDS. It was initially proposed as a cheap alternative to existing treatments, and as such was seen as a good candidate for developing nations, especially those with a high incidence of HIV/AIDS. Given this appeal, and statistical evidence from RCTs supporting the efficacy of the drug, the government of Zimbabwe adopted and endorsed the use of efavirenz in its population in 2015. However, soon after its deployment, reports of patients suffering from adverse neuropsychological events began surfacing, which led to mass withdrawals from the treatment (Nordling 2017).

Notably, mechanistic evidence claiming that this may occur already existed. Park et al. (*forthcoming*) explain that pharmacogenomic research that provided such evidence had been conducted since around 2008.¹⁰ The research showed that a particular gene polymorphism (the CYP2B6*6 allele), that occurs in approximately 20 percent of the Zimbabwean population, led to lower rates of metabolization of efavirenz and that this could cause adverse neuropsychological effects. The evidence fits my characterization of mechanistic evidence in Section 2.2. It provides appropriate-level information (the prevalence of the CYP2B6*6 allele in the Zimbabwean population) about the specific entities (the CYP2B6*6 allele), specific activities (a low rate of efavirenz metabolism), and specific organizational structures of the causal pathway between two variables of interest (diminished response to efavirenz that may lead to neuropsychiatric effects).

Importantly, this evidence was ignored during the approval process of efavirenz in Zimbabwe despite vocal warnings from the researchers involved. Park et al. (*forthcoming*) persuasively argue that a good explanation for the dismissal of this evidence is that it is mechanistic and not the result of RCTs and meta-analyses. They suggest that, as a WHO member state, Zimbabwe is committed to following the organization's recommended guidelines regarding evidence. Since WHO guidance endorsed the statistical approach, the argument goes, it is likely that the Zimbabwean government adopted a dismissive attitude to the existing mechanistic evidence. Not only does this case clarify significant aspects of the debate regarding the statistical and mechanistic approaches, it provides an

⁹ I do not provide an in-depth explication of the Zimbabwe-efavirenz case here; for this, I direct readers to Park et al. (*forthcoming*) for a thoroughly informative exegesis of the events leading to the drug's disastrous rollout in Zimbabwe.

¹⁰ Park et al. cite work by a team led by Zimbabwean geneticist Collen Masimirembwa, particularly Nyakutira et al. (2008). Other work from the same team, including Aklillu et al. (2007), Maimbo et al. (2012), and Dhoro et al. (2015) was dismissed as well. This shows that a large amount of mechanistic evidence that could have been considered existed.

illustrative example of how the Toulmin model I propose can be used to organize existing evidence in a systematic accessible way.

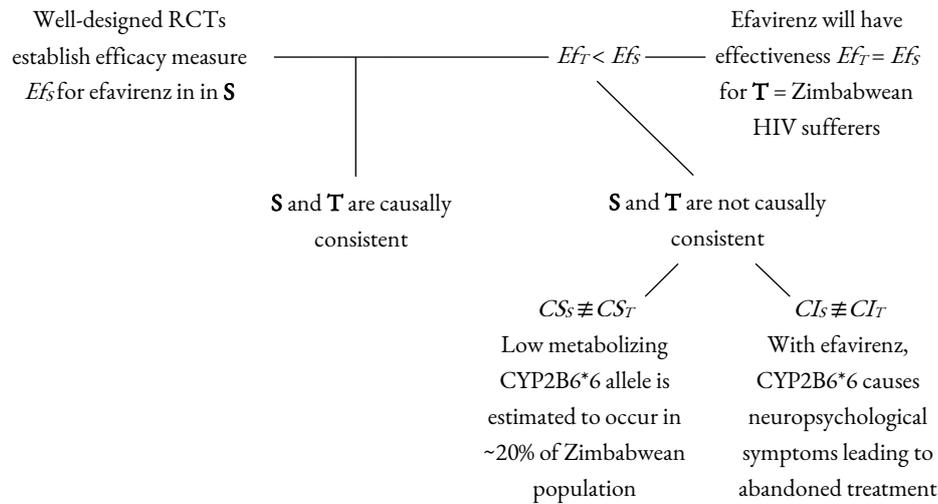


Figure 2.5. Applying the Toulmin model for predicting medical effectiveness of efavirenz in Zimbabwe.

Let us apply the Toulmin model to this case to see how mechanistic evidence can be systematically amalgamated with statistical evidence (Figure 2.5). In this case, the claim was that the effectiveness of efavirenz in the target population of HIV sufferers in Zimbabwe would be similar to that of the efficacy achieved in RCTs sample contexts. No mechanistic evidence was cited in support of the consistency of causal structures, causal enablers, and causal inhibitors across the two contexts. Thus, there is no backing to support the warrant that the predominantly North American populations on which efavirenz was tested would be causally consistent with the target Zimbabwean population. The rebuttal however has support in the form of mechanistic evidence regarding heterogeneity in at least causal structure and causal inhibitors across the two contexts. Available mechanistic evidence shows that around 20 percent of Zimbabweans possess the CYP2B6*6 allele which makes them low metabolizers of efavirenz. Moreover, there is mechanistic evidence that, with efavirenz, the CYP2B6*6 allele is strongly associated with adverse neuropsychological effects which will likely lead to patients abandoning treatment. Therefore, there is a causal inhibitor present in at around 20 percent of the target population. Together, the mechanistic evidence above indicates that the claim should be qualified to recognize that the effectiveness of efavirenz in the Zimbabwean population will likely be substantially lower than it was in the sample populations.

2.6 Why the Toulmin Approach?

In this Section, I respond to a possible objection to my suggestion that the Toulmin model provides a useful way to integrate mechanistic evidence. It might be argued that the Bayesian networks models on offer are better options for mechanistic evidence integration and that the Toulmin model proposed above is superfluous to predicting medical effectiveness. Indeed, Park et al. (forthcoming)

suggest that Pearl and Bareinboim's (2014) *do-calculus* can be used to combine quantitative information about mechanisms with existing statistical evidence. But there are several reasons for my suggestion for a Toulmin model-based alternative. One reason involves recognizing that there are at least two senses of evidence amalgamation. The first is *mathematical amalgamation*—the synthesis of different sets of evidence to produce a numerical measure of the phenomenon of interest. For example, in the context of measuring the effectiveness of medical interventions, the aim of a meta-analysis is to combine statistical results from two or more separate clinical studies to estimate the overall effect of a treatment. Similarly, a Bayesian network aims to accurately determine the probability of an outcome given numerous conditional factors. These methods are thus apt for amalgamating numerical evidence—efficacy measures in the case of meta-analyses and probabilities in the case of Bayesian networks.

The second sense of evidence amalgamation is *systematic amalgamation*—the collection, organization, and evaluation of different sets of evidence such that we know which is relevant to the inferences we draw, and which may should be included in any mathematical amalgamations we might perform. In EBM, this sense of amalgamation commonly takes the form of systematic reviews, which are a performed prior to meta-analysis. A systematic review is the process of identifying, selecting, and evaluating relevant research for a particular research question. The Toulmin model I propose above is similar in this sense in that it involves the organization of existing relevant statistical and mechanistic evidence into a model of reasoning for a particular prediction of medical effectiveness. From this we may see what evidence may be included in Bayesian network models for the given prediction.

Another reason for suggesting the Toulmin model is that relevant evidence is often not translatable into the kind of information used in Bayesian networks. Concrete values for the probabilities in a Bayesian network can sometimes be inaccessible. A Toulmin diagram can help with this by at least sorting out what kind of information should be considered when making predictions regarding a drug's effectiveness in some target context. Although the sophisticated apparatus proposed by others is appealing, the simpler model proposed above may be quite helpful in cases where evidence cannot be straightforwardly included in a Bayesian network model.

One final consideration concerns science communication. Intervention policy options often need to be communicated to policy decision-makers, and decisions conveyed to the public. Often effective decision-making and public trust in those decisions relies on understanding how outcomes from the methods were inferred. However, mathematical amalgamation methods, and thus Bayesian networks, are usually difficult to communicate to non-experts (Douglas 2012). One option to mitigate this drawback is to develop ways in which such methods can be communicated to the uninitiated. Since the Toulmin model I propose can be used to survey which evidence is relevant for Bayesian network methods, it may be seen as one way to communicate the outcomes of such models to policy decision-makers and the public.

2.5 Conclusion

The mechanistic approach to evidence claims that mechanistic evidence—appropriate-level information about the components of the causal pathway between two variables—is important for improving predictions of drug effectiveness. With this approach gaining increasing support in recent years, a deeper understanding of how it differs from the evidential monism of the statistical approach is vital for explaining the different attitudes each has toward evidence in the context of predicting medical effectiveness. Where the statistical approach adopts a dismissive attitude toward mechanistic evidence, the mechanistic approach, whether in its stricter evidential dualist or more amenable evidential pluralist form, embraces an augmentive attitude toward both types of evidence.

The underlying reasons for these disparate attitudes are often said to be based on views regarding evidence quality. The dismissive attitude results from the EBM-centric view that mechanistic evidence is lower in quality than statistical evidence, particularly that gathered from RCTs and meta-analyses. The augmentive attitude, on the other hand, is said to be based on the EBM+ argument that good mechanistic evidence should be considered equal in quality to statistical evidence. However, proponents on both sides agree that mechanistic evidence can be helpful in causal inference. There is thus more to the reasoning behind the dismissive and augmentive attitudes—assumptions regarding contextual similarity. The dismissive attitude is also due to the assumption that there is a low probability that there will be relevant differences between a sample and target context, while the augmentive attitude arises from the assumption that there is a high probability of there being such differences.

The kinds of differences which may be relevant for predicting effectiveness can be captured by the concept of causal consistency. Two contexts are causally consistent to the extent that they are homogenous across three categories: Causal structure, causal enablers, and causal inhibitors. Mechanistic evidence helps in evaluating causal consistency by providing appropriate-level information about which components make up the causal pathway between a drug and its intended outcome, and information about how factors external to the causal structure impact the relation in question.

The view that mechanistic evidence is important for predicting medical effectiveness is unproductive without methods for integrating such evidence with statistical evidence. Based on the Toulmin model of argumentation, I offer a model for the systematic integration of mechanistic evidence with statistical evidence. This model incorporates the concept of causal consistency to organize available mechanistic evidence and evaluate its effect on predictions of effectiveness. The suitability of the model is demonstrated using the case of failed rollout of the HIV treatment, efavirenz, in Zimbabwe. The model I propose is not necessarily an alternative to the Bayesian network methods proposed by others. The latter are good for mathematical amalgamation—integrating quantifiable mechanistic evidence with statistical evidence—whereas the former is appropriate for systematic amalgamation—collecting, organizing, and evaluating mechanistic evidence with statistical evidence.

Chapter 3

P-hacking: Its Costs and When It Is Warranted

3.1 Introduction

P-hacking involves the use of analytic techniques that may lead to exaggerated experimental results. While it is widely condemned, some have suggested that there are some contexts where the practice may be warranted. I have three aims in this Chapter. First, I provide a sorely needed definition of p-hacking. Second, I use philosophical tools from decision theory to articulate the prevalent position on p-hacking and illustrate how serious its effects on obtaining false-positive results can be. And third, I argue against the prevalent position, defending the view that there are scenarios in which p-hacking may be warranted, with a particular focus on non-epistemic judgements.

P-hacking is typically criticized on epistemic grounds. Given that it raises the probability of acquiring false-positive results, the argument goes, there is little warrant for believing the conclusions of studies where p-hacking has occurred (Simmons et al. 2011; Simonsohn et al. 2014). Indeed, such practices have been linked to the replication crises in various fields like psychology and medicine (Munafò et al. 2017; Open Science Collaboration 2015). While these arguments underscore pertinent epistemic features of the case against p-hacking, others mention its practical consequences. Because p-hacking increases false-positive report rates, its regular practice, particularly in fields like medicine, could lead to harmful policies and recommendations based on false findings (Head et al. 2015) and may foster a general distrust of science (Colquhoun 2017).

Such arguments against p-hacking are common in disciplines like economics and psychology, yet few have illustrated the potential severity of its consequences for obtaining false-positive results. Moreover, there seems to be a dearth of philosophical work about how serious the epistemic and practical costs of p-hacking can be. This is an important project; pernicious p-hacking is an ongoing problem, especially in medical research where it could have extremely negative effects on patients in the clinic. I use tools from decision theory, particularly expected utility, to articulate the most common argument against such practices, which I refer to as *the prevalent position* on p-hacking. Framing the argument in this way is useful for at least two reasons. First, it illustrates the enormity of one aspect of the epistemic costs of p-hacking, and second, it describes precise claims of proponents of the prevalent position.

Despite these arguments, some have argued that such practices are, under certain conditions, appropriate (Gelman 2004; Hitzig and Stegenga 2020). I offer my own defense of the thesis that p-hacking can sometimes be warranted. Here, I argue that the prevalent position fails to recognize

other important epistemological and practical features of the debate. I argue that the propriety of analytic choices depends, in part, on there being some reasonable link between the hypothesis and data in question. Such a link may involve some combination of *relevant prior knowledge*, *analogous reasoning*, *plausible mechanism*, or *value-based judgement*. In this Chapter, I focus on the last of these. I apply features of the argument from inductive risk to the case of COVID-19, arguing that value-based judgements can warrant p-hacking (Douglas 2000). I conclude that while p-hacking should generally be considered epistemically pernicious, there is a limited set of cases in which practical concerns may warrant such practices.

Given that there is little in terms of a precise definition of p-hacking in the literature, I begin in Section 3.2 by describing the key features which distinguish it from other related practices. In Section 3.3, I articulate the prevalent position on p-hacking. I begin by outlining a study of progesterone research (Prior et al. 2017) which illustrates the importance of understanding the impacts p-hacking can have on medical inferences. Then, for simplicity, I describe a hypothetical case where a researcher chooses to use methods which amount to p-hacking and acquires statistically significant results. I express this choice as a decision problem and develop an expected utility framework which I use in the rest of the section. Then, by linking previous work on false-positive report rates (Ioannidis 2005) with seminal work on p-hacking (Simmons et al. 2011), I illustrate the potential severity of its epistemic consequences on research results. Incorporating this into the expected utility framework, I describe how the potential epistemic costs of p-hacking are linked to its practical consequences, and articulate the conditions under which proponents of the prevalent position argue p-hacking is harmful. The prevalent position maintains that p-hacking is always epistemically and practically pernicious. I then argue that the prevalent position fails to recognize key epistemic and practical features of p-hacking, suggesting that once these are considered, we can see how p-hacking may be warranted. In Section 3.4, I defend the view that sometimes, analytic choices which amount to p-hacking may be considered appropriate, following which, I outline how practical considerations can be used to warrant practices that many may consider p-hacking. Section 3.5 concludes.

3.2 What is P-hacking?

Strikingly, there is a lack of explicit definition of p-hacking provided in the literature. Without this, assessments of allegedly p-hacked results risk being hindered by needless debates over different interpretations of the practice. Therefore, before addressing the severity of its consequences, it would be worthwhile to have an explicit definition of p-hacking. This involves the conceptual analysis and explication of disparate accounts of the term. I analyze how p-hacking has been characterized in existing work, noting important similarities and inconsistencies between these different descriptions. Then, by outlining its necessary features, I offer a precise definition of p-hacking that I suggest should be used in discussions of the practice.

Exactly what is meant by ‘p-hacking’ is largely taken for granted and the term is used ambiguously and somewhat unreflectively in the literature. Some scholars have offered conflicting

characterizations of p-hacking. For instance, Simonsohn et al. (2013: 534) describe it as the practice of placing statistically insignificant analyses in the file drawer, alluding to intentionally opaque reporting. Yet elsewhere Simonsohn is favorably quoted as stating that “P-hacking...is trying multiple things until you get the desired result” (see Nuzzo 2014: 152). Similarly, Gelman and Loken (2013) see it as an issue concerning multiple comparisons but emphasize, as does Nuzzo (2014), that it may be intentional or unintentional.

Common to most characterizations, then, is the idea that p-hacking involves multiple analyses that may lead to overstated research results. However, this broad description is too vague, making it difficult to differentiate between p-hacking and other related practices, such as *HARKing* (hypothesizing after results are known), *data fishing*, or *circular analysis*. Indeed, several scholars equate p-hacking with other analytic misdeeds and biases. Some prominent biologists, for example, have suggested that p-hacking is synonymous with “inflation bias” or “selective reporting” (Head et al. 2015: 1), and Nuzzo (2014: 152), a leading statistician, has stated that p-hacking is “also known as data-dredging, snooping, fishing, significance-chasing and double-dipping”. A good definition of p-hacking should help differentiate between these practices.

Perhaps what explains the conflation of p-hacking with other related practices is that they all lead to the same bias—*data-dredging bias*. This is a distortion of results arising from probing data in unplanned ways (Erasmus et al. 2020). P-hacking, HARKing, data fishing, and circular analysis all involve the unplanned testing of data and could produce exaggerated results. It is thus tempting to lump these practices together into one general category. While not entirely wrong, there is something to be gained by splitting these practices apart. For one thing, there are nuanced differences between them, which may result in disparate consequences. For instance, where p-hacking, as I illustrate below, can have a large effect on analytic results, HARKing may have less severe consequences.

One way in which p-hacking differs from these related practices, is that it involves probing a data in unplanned ways such that the chance of obtaining the desired results for a *particular prespecified hypothesis* increases.¹¹ Unlike HARKing, which involves analyzing data, obtaining results, and then developing various hypotheses, p-hacking involves having a given hypothesis in mind and examining data in various ways to confirm *that hypothesis*. Likewise, where data fishing involves analyzing different constellations of variables from datasets from multiple sources for putative correlations, p-hacking analyzes different ways in which fixed variables—such as between a specific medication and a particular outcome—may be correlated. As a final comparison, p-hacking is different from the practice of circular analysis in that, where the former analyzes a single dataset in different ways to find a correlation between two variables, the latter analyzes a single dataset, finds a correlation, then uses that result with the same dataset to find other putative correlations.

What makes it possible to test a single prespecified hypothesis in various ways is that there are numerous analytic choices that *must* be made to complete the analysis of any single hypothesis.

¹¹ Note that a “prespecified” hypothesis in this case is not the same as having a pre-analysis protocol.

These choices, often referred to as “researcher degrees of freedom” (RDFs), may be exploited to obtain results which seemingly confirm the hypothesis of interest (Simmons, et al. 2011; Gelman and Loken 2013). RDFs include, but are not limited to, how to define study parameters, how to set exclusion and inclusion criteria, which variable attribute(s) to measure, or what statistical test should be performed on the data. If, for instance, an initial analysis produces results which are close to being statistically significant, then one can make further analytic choices which may produce a statistically significant p-value.¹²

To illustrate, suppose a researcher wants to know whether a new triptan medication is effective for treating migraine pain. The researcher sets up a randomized control trial (RCT) to test the hypothesis that the drug is effective when compared to a placebo. They perform the RCT and gather self-report data about pain relief from participants in both the control and treatment groups, and demographic information for the patients is available. Immediately, there are countless ways in which the difference between the drug and placebo can be compared, all of which are consistent with the collected data. The researcher can measure the difference between men in each group, or the difference between women only. They might measure the effect for certain age ranges, say participants between 31 and 40 years of age. A difference might be observed between participants from a particular socio-economic category (whose classification is open to different measurements), or between those from a particular ethnic background. The researcher might take data about how long it took for participants to no longer feel pain into account and observe a difference. There could very well be no difference for any of these attributes, but one when all participants from each group are considered. And we are yet to even consider the different statistical tests that may be performed on the data.

Overall, there are two important features of RDFs. First, these analytic choices are consistent with the data in question. It is possible that one or more combinations of RDFs may yield statistically significant results that support the hypothesis in question. And second, such choices are unavoidable. We are *required* to settle on some combination of RDFs to test a hypothesis.

The first important feature of p-hacking, then, is that it is the misuse of analytic techniques for a single hypothesis. This misuse of analytic techniques is possible due there being RDFs that must be decided on when testing the hypothesis. The second feature we can derive from the above discussion is that the hypothesis in question is prespecified—the researcher is interested in confirming the hypothesis they started with in the first place. These two features already go some way to differentiating between p-hacking and other related practices.

There is another important feature of p-hacking that helps set it apart. It does not necessarily involve the *intentional* exploitation of RDFs (Gelman and Loken 2013). That is, p-hacking

¹² Note that despite the ‘p’ in ‘p-hacking’ referring to p-values, the practice can occur in studies regardless of what statistical inference tools are being used, including null hypothesis significance testing (NHST), and other frequentist or Bayesian approaches. That is, one may exploit analytic choices in these other approaches to obtain “attractive” results. Given its ubiquity in research and its relative simplicity for my arguments, I focus on NHST in this Chapter.

sometimes arises from a lack of awareness rather than deliberate misconduct. There are times when one's RDFs *appear* entirely appropriate *given what we see in the data*. Imagine a hypothesis is tested by using a (perhaps pre-planned) course of analysis φ , using a unique classical test statistic T , applied to data x , thus yielding an outcome $T(x; \varphi)$. Now consider two scenarios:

(*Scenario 1*): The test is performed n times, with each analysis applied to different constellations of the data, $T(x; \varphi_{1-n}(x))$ and only the best, or most attractive result is reported, thus yielding $T(x; \varphi^{best}(x))$.

(*Scenario 2*): The test is performed based on data x thus yielding $T(x; \varphi(x))$, but *had different data* been observed, a different test would have been performed.

Scenario 1 describes a case of deliberate p-hacking. There are times when researchers intentionally explore a given dataset to attain positive results for a hypothesis of interest, although, as expected, this often occurs with other misconduct like the direct manipulation of data. For example, consider the triptan case above. The researcher might test several constellations of RDFs and find a statistically significant result for women between the ages of 31 and 40, but not for any other combinations. *Scenario 2*, on the other hand, involves the unintentional exploitation of RDFs. Here, φ is a function of the data observed. In such a scenario, it may be that the human tendency to unconsciously interpret ambiguity in a self-serving fashion leads to these analytic choices, particularly when these decisions are not *strictly* prespecified. To illustrate, say the triptan researcher initially tests for a difference between women in the control and treatment groups and observes a near-significant p-value. Based on this, they then analyze women between 31 and 40 years of age and find a statistically significant difference. This result seems reasonable given the previous near-significant result, however, she has unconsciously exploited the RDFs for the hypothesis. Moreover, had the results from her initial analysis not been near-significant, she would not have performed the age-based analysis. In other words, many researchers assume they are properly analyzing the data when, in fact, they are unintentionally p-hacking. This may be a more common occurrence than the first scenario, but it should be noted that both *Scenario 1* and *Scenario 2* involve the exploitation of RDFs and both should be considered p-hacking.

We are now well-positioned to provide a definition of p-hacking that should be used in practice.

P-hacking is the intentional or unintentional exploitation of researcher degrees of freedom, which may lead to exaggerated results regarding a single, prespecified hypothesis of interest.

This definition helps to more reliably identify cases of p-hacking and differentiate between its practice, and other analytic offenses.

Before focusing on the possible consequences of p-hacking, it is worth addressing a potential concern about my definition. It may be argued that p-hacking is not just a question of analytic practices, but also one of scientific integrity. Put another way, it comes down to whether one is

transparent about their analytic procedures. It therefore may be argued that another necessary feature of p-hacking is that it involves some *opaque*, *selective*, or *inaccurate* reporting of analytic procedures or results. Head et al. (2015: 1), for instance, claim that p-hacking “is the misreporting of results”. Generally, transparently reported courses of analysis are considered acceptable because knowing what analyses took place in a study can help calibrate one’s confidence in the results (Ioannidis 2008).

However, failure to accurately report one’s analytic procedures is neither necessary nor sufficient for p-hacking. A researcher can p-hack without being any more opaque, inaccurate, or selective in their reporting than any non-p-hacked study. For instance, in the triptan example, the researcher may have exploited RDFs and listed all their assumptions, analyses, and accurately reported the results they acquired. That they were transparent in their reporting may justifiably lower our confidence in their results, but it does not change the fact that they exploited RDFs and acquired possibly inflated support for their hypothesis. Besides, the norms of evidence reporting do not require that one disclose *every* assumption made and *every* analysis performed. A researcher can selectively or inaccurately report their methods or results without having p-hacked. The triptan researcher may not have exploited RDFs at all and still not have accurately reported their analyses. This may explain why some have interpreted the case against p-hacking as one of scientific integrity and not the typical failure to understand its consequences. In the next section, I turn the question of how serious those consequences can be.

3.3 The Impact of P-hacking on Results

Arguments against p-hacking operate on the logic that the chance of discovering a statistically significant result from numerous analyses is higher than if just one analysis is performed (Ioannidis 2005, 2008; Simmons et al. 2011; Nuzzo 2013). Although compelling, this reasoning does not quite capture the extent to which p-hacking may affect the probability that a given result is false. In this section, I provide a detailed formulation of the prevalent position on p-hacking, which is consistent with intuitions about multiple analyses, and which illustrates the magnitude of its epistemic consequences while linking practical concerns. I begin by outlining a real case of medical research where p-hacking is suspected of producing exaggerated results. This case illustrates why we should worry about p-hacking in the first place.

3.3.1 Why P-hacking Matters

Some of the characteristics of p-hacking make it difficult to detect in published results. One proposed method, *p-curve analysis*, aims at detecting selective examines the distribution of statistically significant p-values for a set of results (Simonsohn et al. 2014, 2015). However, such methods require vast amounts of data from many studies for effective detection, and there is little evidence for the validity of the method. What is more, in non-randomized studies, confounding may make tests for p-hacking inaccurate (Bruns and Ioannidis, 2016). An in-depth analysis of such techniques would constitute a project on its own, and so these discussions are beyond the scope of

this particular Chapter.¹³ It should be noted, however, that the challenges to detecting p-hacking make its incidence difficult to gauge and its effects hard to correct for in practice. Even so, whatever impacts there are will directly influence medical inferences. For example, if the evidence we have for an efficacy claim consists of p-hacked results, and p-hacking increases the chance that those results are false, then our predictions will be distorted from the start. An example will make this point clearer.

Progestogens, a treatment aimed at mitigating pregnancy loss, have been found to be effective in several trials and meta-analyses. However, more recent trials have concluded they are ineffective. These conflicting findings led to a recent systematic review and meta-analysis which included only the results of primary outcomes of preregistered trials (Prior et al. 2017). Researchers scoured professionally recognized trial registries finding 194 relevant studies, 29 of which were systematic reviews with meta-analyses; these were grouped together into a Total Study Group (TSG). Altogether, the TSG reported the results of 93 RCTs.

Of the 93 RCTs used in the TSG meta-analyses, 22 were judged as having a low susceptibility to p-hacking because they reported pre-registered primary outcomes. Of these 22 trials, only *one* produced statistically significant results favoring progestogens over placebo. A meta-analysis of these trials found that there was no evidence that progestogens prevented pregnancy loss. In contrast, 19 out of 29 previous meta-analyses found that progestogens were effective. Prior et. al. (2017) suggest that the difference is likely due to the inclusion of p-hacked studies in the previous meta-analyses. While it may be the case that there are many other biases at play here, this is unlikely. If, for example, this difference were the result of publication bias, then, given the 21:1 ratio of non-significant to significant results in pre-registered trials, we should expect that there are over 1,000 unpublished studies which found no statistically significant effect.

This example shows that if (enough) p-hacking occurs, it may lead to questionable efficacy claims, which meta-analyses may not be able to overcome. This, added to arguments for the ubiquity of p-hacking, make it an important issue, not just for predictions of medical effectiveness, but for other medical inferences too. These include inferences about clinical resource allocation policies derived from studies comparing the outcomes of different approaches to care, diagnostics based on studies aimed at reducing error rates, and prognostics based on studies on disease risk factors.

Indeed, there are arguments for the ubiquity of p-hacking in published research. Some claim that if all research were analyzed properly and published accordingly, then we should see roughly the same proportion of p-values just above and just below $p = 0.05$ in the published literature. But there are an unexpectedly large number of studies which report a p-value just below this threshold in medical journals (Jager and Leek, 2014; Albarqouni, et al., 2017; Perneger and Combescure, 2017; Ioannidis, 2019).

¹³ Indeed, a prior question is how p-hacking should be defined. This may be particularly important if different techniques used to detect p-hacking utilize definitions that are inconsistent with one another.

Despite the perceived pervasiveness of p-hacking, there is debate about the severity of its epistemic and practical consequences. Head et. al. (2015) conclude that while p-hacking is rife, it is unlikely to have substantial effects on meta-analyses. This, they argue, is because only smaller studies are vulnerable to p-hacking, and these have negligible effects on the results of meta-analyses. However, their argument assumes that meta-analyses will include large studies with smaller ones. One problem here is that meta-analyses typically consist of small studies. Indeed, there is reason to think that p-hacking may affect downstream medical inferences. It would thus be good to have an idea of just how much impact p-hacking can have on efficacy claims. By illustrating how serious the impacts of p-hacking on the probability of obtaining false-positive results can be, I demonstrate that most researchers may be greatly underestimating the consequences of such practices.

3.3.2 A Case of P-hacking

Consider the following hypothetical study:

(H-trial): A researcher performs a well-designed RCT to test the efficacy of a drug for reducing the risk of heart attack. She chooses to measure the effect of the drug on two variables, blood pressure (BP) and low-density lipoprotein cholesterol (LDL), analyzing each individually and then by combining them. She finds that there is no significant difference in either BP or its combination with LDL between the drug and control groups, but she notices that the result is close to $p < 0.05$ for LDL, so she decides to collect more data. Further analysis reveals that she is closer to obtaining statistical significance for the LDL variable, so she decides to control for the interaction of gender with the treatment and detects an ostensibly better effect in males. Convinced she is onto something, she analyzes the data again by running low effect, medium effect, and high effect conditions. By dropping the low effect condition, she obtains the statistically significant result she knew was there all along: the drug is effective for reducing LDL cholesterol in males and thus, she infers, for reducing risk of heart attack.¹⁴

The practices in *H-trial* amount to p-hacking. The researcher exploits RDFs through redefining study parameters, dropping and combining different variables, and performing multiple analyses.

The researcher's choice to p-hack in *H-trial* can be described using tools from decision theory. The points at which the researcher decides to manipulate and reanalyze the data can be expressed as acts, either *p-hack* or $\neg p-hack$. Her results will either be positive or negative, viz. either exhibit a statistically significant correlation or not, and they will either be true or false. One of the following states will thus occur: true-positive results (r_1), false-positive results (r_2), true-negative results (r_3), or false-negative results (r_4). Further, these states will occur with some probability $P(r_i)$ and, since we

¹⁴ This example is formulated by combining parts of an example given by Hitzig and Stegenga (2020) and statistical methods simulated by Simmons et al. (2011).

are dealing with one set of results from the study, the sum of these probabilities will be unity: $P(r_1) + P(r_2) + P(r_3) + P(r_4) = 1$. Lastly, utilities can be attached to each outcome of the researcher's chosen act over the possible states (see Table 3.1).

	r_1	r_2	r_3	r_4
p-hack	u_1	u_2	u_3	u_4
\neg p-hack	u_1	u_2	u_3	u_4

Table 3.1.

Without knowing that p-hacking occurred, when presented with positive results we would arguably believe that they are true, and if they are in fact true, then we would put them to good practical use. Thus, assume we believe that the consequences, and thus the utilities, of obtaining true-positive results will be the same regardless of whether p-hacking occurred or not. This goes for all other states too, hence:

$$\begin{aligned}
 u(p\text{-hack}\cdot r_1) &= u(\neg p\text{-hack}\cdot r_1) = u_1 \\
 u(p\text{-hack}\cdot r_2) &= u(\neg p\text{-hack}\cdot r_2) = u_2 \\
 u(p\text{-hack}\cdot r_3) &= u(\neg p\text{-hack}\cdot r_3) = u_3 \\
 u(p\text{-hack}\cdot r_4) &= u(\neg p\text{-hack}\cdot r_4) = u_4
 \end{aligned}$$

With all these elements in place, we can articulate the argument from proponents of the prevalent position and examine the severity of the consequences of p-hacking on false-positive report rates using an expected utility approach.

3.3.3 *The Prevalent Position and the Consequences of P-hacking*

Broadly, proponents of the prevalent position claim that researchers should not p-hack because it is epistemically detrimental, and thus could have practically harmful consequences. This suggests two elements to the argument against p-hacking, one epistemic and one practical. By way of caveat for this section, I am not arguing that we should agree with the prevalent position, but rather articulating what I see as the typical argument against p-hacking. Ultimately, I will argue that the prevalent position neglects key features of the discussion around p-hacking, and that, once these are considered, there are scenarios where p-hacking can be warranted (Section 3.4).

Besides the influential work of Ioannidis (2005) and Simmons et al. (2011), which I discuss in detail here, the literature is replete with admonitions of p-hacking for its epistemic effects. These criticisms have come from leading scholars, academic societies, and governmental organizations. The American Statistical Association (ASA), for example, released a statement affirming that p-hacking “leads to a spurious excess of statistically significant results in the published literature and should be vigorously avoided” (Wasserstein and Lazar 2016: 132). Indeed, Simonsohn et al. (2014: 534) state that “p-hacking can allow researchers to get most studies to reveal significant relationships between

truly unrelated variables”. This view is also exemplified in calls for journals to require that studies be preregistered to qualify for publication, which many argue helps prevent p-hacking (Simmons et al. 2017; Hawkes 2018). The American Psychological Association, for instance, notes that one promising feature of preregistered trials is that they are less susceptible to p-hacking (Gonzales and Cunningham 2015). Additionally, the European Union recently released a report titled ‘As Predicted: Preventing p-hacking’ stating that there “is increased awareness about the benefits of preregistration to combat p-hacking” (European Commission, Directorate-General for Research and Innovation 2019: 11).

Many have noted that the epistemic problems caused by p-hacking could potentially have dire practical consequences. One broad consequence is encouraging a general distrust of science (Colquhoun 2017). More concretely, some highlight issues that may arise from developing of policy based on incorrect inferences which result from p-hacking (Head et al. 2015). For instance, despite the increased probability of the false-positive results in *H-trial*, the drug might, on account of the significant result, be deemed safe and effective by administrative bodies. This could result in the it being prescribed to patients with high cholesterol, which may cause harm through possible side effects, costs, or ineffectiveness. Put another way, the prevalent position argues that because its epistemic problems could potentially have harmful practical consequences, not p-hacking is preferable to p-hacking. Unpacking this prevalent position against p-hacking will reveal specific features of the argument that have been overlooked, and which help us recognize the circumstances under which p-hacking may in fact be warranted.

A useful way of understanding the conditions set out in the prevalent position is to articulate it in terms of expected utility (see Appendix 1). Not p-hacking is preferable to p-hacking just in case the expected utility of not p-hacking is greater than the expected utility of p-hacking:

$$\text{not p-hacking} \succ \text{p-hacking} \text{ iff } EU(\neg p\text{-hack}) > EU(p\text{-hack})$$

The right side of this biconditional can be articulated more precisely as:

$$\sum_{k=1}^4 u_i \cdot P(r_k | \neg p\text{-hack}) > \sum_{k=1}^4 u_i \cdot P(r_k | p\text{-hack}) \quad (3.1)$$

Note that one cannot appeal to differences in utilities under each state to account for any difference in the expected utility of each act. These, as previously explained, are equal given the state in question. However, the inequality can be defended by appealing to differences in the probabilities of each state occurring. Here, we are interested in the probability of a state r_k obtaining, given that either *p-hack* or $\neg p\text{-hack}$ occurs. This is the prevalent position’s epistemic argument against p-hacking—we should always choose not to p-hack because if we do and obtain positive results, then the fact that p-hacking occurred makes it more likely that those results are false. In other words, proponents of the prevalent position argue that the probability that the result is a false-positive given that p-hacking occurred is higher than it would be if the results were not p-hacked:

$$P(r_2|p\text{-hack}) > P(r_2|\neg p\text{-hack}) \quad (3.2)$$

This formalization represents the typical epistemic stance on p-hacking. It relies on the intuition that multiple analyses increase the chance of getting desired results. Imagine you were given the following two options:

- (a) I will toss a coin. If it lands heads, I will pay you £100.
- (b) I will toss a coin twenty times. If it lands heads on any toss, I will pay you £100.

Most people would correctly choose option (b) because they realize that, despite the probability of the coin landing heads in each toss is 0.5, there is a higher probability of it landing heads at least once in twenty tosses than there is of it when tossed once. The reasoning can be applied to p-hacking. Since p-hacking involves performing multiple analyses on a set of data, there is a higher chance of at least one of these analyses producing a statistically significant result than there is of a single analysis of the same data producing a statistically significant result.

Still, it might be argued that the intuition alone does not give us reason to doubt the validity of a statistically significant result. To understand this, consider the following: Say that we accept the typical α -value cut-off of 0.05. This essentially means that in a study of the relationship between two variables there is a 5% chance of finding a significant result even if there is no true relationship between the two variables. Now say that we perform the 20 separate analyses on the data and all but 1 of these produces non-significant results. Knowing that 19 out of 20 results were non-significant throws serious doubt on the 1 significant finding. The more separate analyses performed on a set of data, the more likely it becomes that a significant result will be found and if all the analyses bar one result in non-significant findings, the more likely it is that the one significant result is a false-positive. This is akin to a company in the business of producing parachutes securing contracts by only showing clients the 1 instance in 20 in which their parachutes in fact bring someone safely to ground. If the client knows about the 19 failed tests, they will very likely not buy any parachutes from the company.

Returning to *H-trial*, the researcher has performed numerous analyses and not acquired a significant p-value in all but one. Given that all but one of her analyses produce non-significant p-values, it is less certain that the one significant result is true. Of course, the researcher has only performed four analyses on the data, and so, one might be inclined to think that this is not as questionable as performing 20 analyses. This conclusion would be too hasty, however. In what follows, I illustrate the extent to which p-hacking affects the probability of obtaining false-positives. To do this, and to show more precisely how (3.2) holds, I appeal to two discussions. In the first, Ioannidis (2005) famously illustrates how to get an idea of the likelihood that a published positive result is false. For simplicity, take the following example:¹⁵ Given a set of 1000 testable hypotheses, assume that 100 of them are true—the proportion of true hypotheses is 10%: $\pi = 0.1$. When testing

¹⁵ This example is taken from Forstmeier et al. (2017), who neatly outline the essential point of Ioannidis' argument.

the other 900 false hypotheses, we typically allow a 5% false-positive rate: $\alpha = 0.05$. Thus, we will obtain 45 false-positive results. Further, we will reject a number of the 100 true hypotheses because the data does not provide significant support. This false-negative rate β is determined by our sample size and effect size. The larger the dataset, the lower the risk of failing to obtain significance. Hence, a large dataset means that the study has high statistical power: $1 - \beta$. Given the large sample size in the example, we have a power of 80%, and thus there is a 20% false-negative rate: $\beta = 0.2$. Importantly, when considering the set of positive outcomes where a hypothesis has evidentiary support from the data, a larger proportion than we might expect will be false. The fraction of positive results which are false is the false-positive report probability (FPRP):

$$\text{FPRP} = \frac{\alpha(1 - \pi)}{\alpha(1 - \pi) + (1 - \beta)\pi} \quad (3.3)$$

By substituting the values from our example into (3.3), we get a probability of 0.36 that an obtained positive result is false. Ioannidis thus demonstrates that stipulating that $\alpha = 0.05$ does not actually entail that only 5% of positive results are false.¹⁶ Note that in Ioannidis' argument, (3.3) gives us an idea of the probability that a positive result is false *without* taking p-hacking into account: $P(r_2 | \neg p\text{-hack}) = 0.36$.

Whereas Ioannidis provides a way to determine the FPRP where p-hacking has not occurred, a second account, by Simmons et al. (2011), can be used to illustrate how it is affected by p-hacking. In other words, with this account we can determine the probability of obtaining false-positive results given p-hacking: $P(r_2 | p\text{-hack})$. This discussion highlights how a lack of constraints on RDFs allows one to obtain statistically significant results from just about any data. Computer simulations show that just four combined decisions to use an analysis that is better for producing lower p -values, much like the choices made in *H-trial*, increases α from 0.05 to 0.607. This means that making repeated choices to p-hack substantially increases probability of producing statistically significant results when no real effect exists.

This illustrates how *massive* the effect of p-hacking on the FPRP can be. Many take the increased α -value to mean that if p-hacking occurs, up to 61% of obtained positive results are false (see Simmons et al. 2011). But this, despite seeming like a huge effect, is an underestimation. To see why, we just need to substitute the simulation findings into the example of 1000 hypotheses from before. The effect this has on the FPRP is enormous. With $\alpha = 0.607$ the FPRP increases from 0.36 to a colossal 0.87. Put another way, if the data were repeatedly p-hacked, then there is an 87% chance that a given positive result is false: $P(r_2 | p\text{-hack}) = 0.87$.

Taking Simmons et al. and Ioannidis' accounts together illustrates just how epistemically problematic p-hacking can be—such practices can increase the probability that a given positive result

¹⁶ Things get worse when we consider that many published studies have lower power due to small sample sizes. If a study has a statistical power of 25%, then the probability of a positive result being false goes up to 0.66.

is false from 0.36 to 0.87.¹⁷ Second, in doing so, it vindicates (3.2)—it shows that $P(r_2|p\text{-hack}) > P(r_2|\neg p\text{-hack})$. Finally, it illustrates the following epistemic condition given by proponents of the prevalent position (PP_E): P-hacking increases the probability that an obtained positive result is false.

Of course, PP_E does not, on its own, entail that (3.1) holds. Despite not being able to appeal to differences in utilities *under* each state, (3.1) relies on differences in utilities *across* states, viz. differences such that (3.1) results in a higher expected utility for not p-hacking than for p-hacking.

It is reasonable to assign utilities based on what we hope the research will achieve. Since having a true effect will help those with high LDL cholesterol and a false effect will not, a true-positive result in *H-trial* will have better consequences than a false-positive one, $u_1 > u_2$. Having a true-negative and a false-negative in *H-trial* will arguably result in the same utility since there will be no consequences given that no policy will be developed thus, $u_3 = u_4$. And since, in this case developing no policy is better than developing policy based on false findings, obtaining a true-negative result is better than obtaining a false-positive one, $u_3 > u_2$. Finally, a false-negative result would be worse than a true-positive one, $u_1 > u_4$, because we do not develop an effective policy when we may have. In line with this, suppose that $u_1 = 100$; $u_3 = u_4 = 50$; and $u_2 = 25$.

Now we can substitute these utilities and the probabilities from before into (3.1). Assuming a positive result is obtained without any p-hacking occurring, $P(r_1|\neg p\text{-hack}) = 0.64$ and $P(r_2|\neg p\text{-hack}) = 0.36$. Thus, the expected utility of not p-hacking is calculated as:

$$EU(\neg p\text{-hack}) = (0.64 \times 100) + (0.36 \times 25) = 73$$

Now, assuming a researcher decides to perform four consecutive analyses like those in *H-trial*, $P(r_1|p\text{-hack}) = 0.13$ and $P(r_2|p\text{-hack}) = 0.87$. Thus, the expected utility for p-hacking the data is:

$$EU(p\text{-hack}) = (0.13 \times 100) + (0.87 \times 25) = 34.75$$

This shows that, at least in this case, the practical consequences of p-hacking are worse than those of not p-hacking. Moreover, those who argue for the prevalent position suggest that this is true for all cases in which PP_E holds and where $u_1 > u_2$. This latter conjunct is the practical condition given by proponents of the prevalent position (PP_U): Obtaining true-positive results entails more beneficial outcomes than false-positive results.

The argument from expected utility I have outlined above illustrates the potential severity of p-hacking on false-positive report rates, and how p-hacking can lead to harmful practical consequences. Additionally, it outlines two central claims of the prevalent position: Proponents of the prevalent position hold that p-hacking should always be avoided because it increases the probability that obtained positive results are false, and because false-positive results are always less desirable than true-positive results. However, the prevalent position faces some immediate

¹⁷ Simmons et al. (2011) state that the results of their simulations may actually be conservative. Their simulations do not consider other common practices in statistical research which may increase α -values. Furthermore, they illustrate that with just one p-hacking decision, the α -value increases from 0.05 to 0.08 which results in an FPRP of 0.47.

challenges. In the next subsection, I outline some of these and then, in Section 3.4, expand on the argument that p-hacking can, in fact, sometimes be warranted.

3.3.4 Immediate Challenges to the Prevalent Position

There are striking issues facing the typical argument for the prevalent position. The first is that it is not so clear that the p-hacking is always entirely epistemically harmful. To see why, we need to consider the *overall probability of error* due to p-hacking.¹⁸ Consider a straightforward example in which the following probabilities occur when one does not p-hack. The probability of accepting hypothesis H and H is true is $r_1 = 0.2$, the probability of accepting H and H is false is $r_2 = 0.05$, the probability of rejecting H and H is false is $r_3 = 0.35$, and the probability of rejecting H and H is true is $r_4 = 0.4$ (see Table 3.2).

	H is True	H is False
Accept H	0.2	0.05
Reject H	0.4	0.35

Table 3.2.

Now suppose that one p-hacks and this results in the following probabilities, which are consistent with there being an increase in the probability of obtaining false-positive results. The probability of accepting hypothesis H and H is true is $r_1 = 0.555$, the probability of accepting H and H is false is $r_2 = 0.3$, the probability of rejecting H and H is false is $r_3 = 0.1$, and the probability of rejecting H and H is true is $r_4 = 0.005$ (see Table 3.3).

	H is True	H is False
Accept H	0.555	0.3
Reject H	0.005	0.1

Table 3.3.

Notice that the overall probability of error in the non-p-hacked experiment represented in Table 3.2 is $P(r_2 \vee r_4) = 0.45$, while the overall probability of error in the p-hacked experiment in Table 3.3 is $P(r_2 \vee r_4) = 0.305$. Therefore, we have a case where, when the overall probability of error is considered, p-hacking is less epistemically harmful than not p-hacking. Notice too that in Table 3.3, the probability that we accept a true hypothesis has risen from 0.2 to 0.555. While p-hacking raises the probability of obtaining false-positive results, it also raises the probability of obtaining true-positive results. Indeed, this is one of the key insights of Hitzig and Stegenga's (2021) Bayesian analysis of the epistemic harms of p-hacking.

¹⁸ I would like to thank Daniel Steel for his insightful remarks, which contributed greatly to formulating this point.

The above considerations reveal a further challenge to the prevalent position. The prevalent position focuses on the probabilities and consequences of accepting false-positive results at the expense of the probabilities and consequences of rejecting false-negative results. This is evident in claims from vocal critics of p-hacking cited previously. To reiterate, Simonsohn et al. emphasize that p-hacking allows researchers to obtain spurious “significant relationships between truly unrelated variables” (2014: 534), while the ASA claims that p-hacking should be avoided because leads to an “excess of statistically significant results in the published literature” (Wasserstein and Lazar 2016: 132). Common to these criticisms is an overemphasis on the negative consequences of false-positive results, and neglect of considerations about the consequences of rejecting results that are in fact true. Note that the prevalent position’s conditions, PP_E and PP_U , do not constitute necessary and sufficient conditions for p-hacking being worse than not p-hacking. Consider a simple toy example where $P(r_1|p\text{-hack}) = 0.3$, $P(r_2|p\text{-hack}) = 0.25$, $P(r_3|p\text{-hack}) = 0.25$, and $P(r_4|p\text{-hack}) = 0.2$. Further, say that $P(r_1|\neg p\text{-hack}) = 0.25$, $P(r_2|\neg p\text{-hack}) = 0.2$, $P(r_3|\neg p\text{-hack}) = 0.3$, and $P(r_4|\neg p\text{-hack}) = 0.25$. Note that these probabilities satisfy PP_E , $P(r_2|p\text{-hack}) > P(r_2|\neg p\text{-hack})$. the utility of a true-positive (u_1) is 100, the utility of a false-positive (u_2) is -10 , a true-negative (u_3) 10, and a false-negative (u_4) is -100 . Note that these utilities satisfy PP_U : $u_1 > u_2$. These result in the following results for the expected utilities for p-hacking and not p-hacking:

$$EU(p\text{-hack}) = (100 \times 0.3) + (-10 \times 0.25) + (10 \times 0.25) + (-100 \times 0.2) = 10$$

$$EU(\neg p\text{-hack}) = (100 \times 0.25) + (-10 \times 0.2) + (10 \times 0.3) + (-100 \times 0.25) = 1$$

So, while the upshot of the prevalent position *seems* compelling, there are toy examples in which it does not hold. For the prevalent position to hold, not only do PP_E and PP_U need to hold, but, at the very least, the probabilities attached to all possible outcomes and the consequences of rejecting false-negatives would always need to be such that p-hacking results in lower utility than not p-hacking. Put another way, PP_E and PP_U are insufficient for the biconditional represented by 3.1 to hold. It may be these conditions show why p-hacking in *H-trial* is epistemically and practically harmful, but ultimately this is because rejecting a true association will not result extremely negative utility since there are other treatments for high cholesterol on the market.

It remains to be seen if there are legitimate examples of cases like those in the toy examples provided above. In the next section, I argue that p-hacking can sometimes be warranted by appealing to features of the argument from inductive risk and provide an example of a case where the consequences of rejecting a false-negative result may have severe enough consequences that p-hacking is warranted.

3.4 Warranting P-hacking

While many maintain that p-hacking is both epistemically and practically harmful, there are some who argue that it is sometimes warranted. Gelman (2004), for example, advocates a two-stage analysis where the first is an exploratory stage in which p-hacking (and other related practices) are acceptable, and the second is a confirmatory stage where the findings from the first stage are

(dis)confirmed. In this section, I argue for the view that the choices about which observed data and methods to use when testing a hypothesis can be appropriately made and, following this, use elements from the argument from inductive risk to show how p-hacking may be warranted.

3.4.1 *The Propriety of Analytic Choices*

As mentioned, RDFs are unavoidable in hypothesis testing (Gelman and Loken 2013). In *H-trial*, for instance, the researcher *had* to make a decision about whether to examine the relationship between the drug and either BP, LDL, or both because the results from all three have something to say about the overarching hypothesis. This decision, in particular, highlights an important detail about analytic decisions: Some choices are more *appropriate* than others. For example, choosing to examine the relationship between the drug and, say, BP is more appropriate than choosing to analyze the relationship between the drug and average body temperature. On the one hand, since there is no established relation between one's average body temperature over time and the risk of heart attack, examining this relation would be questionable. The researcher's decision to examine the relationship between the drug and BP, on the other hand, is more appropriate given that there is well-established background knowledge regarding the relation between persistent high BP and the risk of heart attack. The choice is made, in this case, based on previous evidence and background knowledge.

This reasoning can be similarly applied at the level of choices regarding RDFs. Recall that after getting the results from the initial analyses on BP, LDL, and the combination of the two in *H-trial*, the drug has an ostensibly better effect on LDL. Imagine now that in addition to this, the chemical structure of the drug in question is similar to other pharmaceuticals which reduce LDL. This, taken with the ostensibly better effect add to the propriety of the researcher's choice to collect more LDL data, particularly when compared to a case in which there were just an ostensibly better effect. This suggests that there are scenarios in which there is good reason to think that a particular combination of RDFs is appropriate.

Recall *Scenario 1* and *Scenario 2* from Section 3.2. Gelman and Loken (2013) distinguish between cases of blatant p-hacking—where RDFs are intentionally exploited to obtain an attractive result—and cases of unintentional p-hacking—where the test is chosen based on observed data. In these latter cases, analytic choices *appear* entirely appropriate to a researcher *given what they see in the data*. What I am proposing here is the existence of scenarios like the following:

(*Scenario 3*): The test is performed based on data x and background knowledge b which links the observed data to the hypothesis of interest, thus yielding $T(x; \varphi(x; b))$, but *had different data* been observed, no other test would have been performed. Here, φ is a function of the data observed and the background knowledge linking the hypothesis to the data.

Cases like *Scenario 3* are appropriate because they include information *beyond the data in question*, which links the observed data with the hypothesis of interest. This information increases one's confidence in that the analytic choice is warranted.

The information underlying the propriety of a particular analytic choice in such scenarios will differ from case to case but may include some combination of *prior knowledge regarding variables*, *a proposed plausible mechanism*, or as above, *reasoning from analogous theories or cases*. Put another way, just because a testable hypothesis can be (dis)confirmed by multiple constellations of evidence does not always mean that the decisions regarding how to interpret and analyze observed data are always unfounded.

Besides those listed above, another possible set of information for making decisions about RDFs may come from non-epistemic considerations of inductive risk. By applying the argument from inductive risk (AIR) (Rudner 1953), particularly Douglas' (2000) extension of the argument, to what is revealed in the expected utility approach, I argue that there are cases in which p-hacking may be warranted on practical grounds.

3.4.2 *Warranting P-hacking using Non-Epistemic Judgments*

Another way to argue against the prevalent position is to object to PP_U . Recall, this is the condition states that obtaining true-positive results entails more beneficial outcomes than false-positive results. So, one would need to demonstrate that there are hypotheses for which the expected utility of false-positives is higher than for true-positives. Here, since $u_2 > u_1$ entails that p-hacking will bring about a higher expected utility than not p-hacking, we acquire the strange result that p-hacking is preferable to not p-hacking. However, this argument is implausible since practicing researchers would seldom think that false-positive results are more beneficial than true-positive ones. Of course, p-hacking cannot be warranted based on the utility of obtaining false-negative results.

Another question to ask, however, is whether a researcher's exploitation of RDFs—and thus decision to p-hack—could ever be warranted based on non-epistemic judgements of the consequences of being wrong. This involves appealing to an extended version of the AIR, which claims that non-epistemic judgements of inductive risk are not restricted to setting significance thresholds but are also present *within the research process itself*. Douglas (2000) convincingly illustrates that there are at least three areas within a study where researchers could make errors which may have non-epistemic consequences: the characterization of data; the choice of model; and the interpretation of results. RDFs is another such area, fraught with inductive risk, where non-epistemic values play a role. This, after all, is another way of characterizing the prevalent position: Because wrongly deciding to use a particular combination of RDFs can have serious consequences, we should avoid using them. However, I argue here that the extended AIR provides a more nuanced approach to such choices, one more amenable to situations where such practices can be valuable.

The prudent researcher knows that making analytic decisions which amount to p-hacking increases the probability that an obtained positive result is false. That said, it is plausible that if the consequences of *wrongly rejecting* a hypothesis are serious enough, then such decisions can be made

despite this increased risk. Since wrongly rejecting the hypothesis in *H-trial* would not necessarily entail that there are no potentially effective treatments available, such considerations may not be a factor for the researcher. Yet, there are cases in which such considerations are extremely important.

One such set of cases are those which require urgent discovery. For example, in the context of a sudden pandemic, such as the that of COVID-19, researchers may be warranted, to a certain extent, in combining RDFs (in a way that amounts to p-hacking) in order to discover potentially effective treatments. It seems that just this sort of reasoning has contributed to calls for the repurposing of existing drugs for COVID-19 research. Given the “high infectivity” of the virus, some scholars have stated that “efficacious treatment, prevention of spread, and vaccine development are global imperatives” (Farne et al. 2020: 1186). The urgency of discovering effective treatments and vaccines for such a pandemic could make accepting false-positives more tolerable.

Of course, possible candidates for treatments and vaccines should and do undergo further scrutiny before being approved, but in cases where urgency is a factor, and the consequences of not finding an effective intervention fast are potentially very serious, it could be argued that such practices may be just what is needed. This is because, while p-hacking raises the probability of uncovering false-positives, it also increases the chances of discovering true-positive results. To be sure, as the prevalent position illustrates, there will be a high probability that such positive results represent spurious correlations. But in cases where the need for discovery is pressing, such results can be examined further and evaluated in terms of background knowledge. And if confirmation increases, these results may turn out to be reliable. What is more, if the consequences of wrongly rejecting a potential candidate treatment may lead to extensive loss of lives, then polity may be somewhat guided by research that engaged in practices which amount to p-hacking.

There is some evidence that these kinds of research practices took place in the early stages of COVID-19 research. To illustrate, consider the case of *dexamethasone*. As of November 2020, this corticosteroid is strongly recommended by the World Health Organization (WHO) for “the treatment of patients with severe and critical COVID-19” (WHO Living Guidance 2020). This recommendation, introduced in late July 2020, was initially based on *preliminary reports* from the RECOVERY trial (The RECOVERY Collaborate Group 2020; Horby et al. 2020). The goal of the trial is to assess the efficacy of various readily available treatments against COVID-19. These treatments, of which one is dexamethasone, have previously been tested for other uses, so, for the most part, their safety has been established (Farne et al. 2020). Researchers in the trial tested the efficacy of dexamethasone against severe COVID-19 for at least four difference subsets of a single group of participants.¹⁹ The results showed that the use of dexamethasone significantly reduced mortality in patients receiving machine ventilation and those receiving supplemental oxygen without ventilation. Results also showed a significant overall reduction in mortality in patients taking the drug, but no benefit for those not receiving any respiratory support.

¹⁹ They may have tested the drug’s efficacy for other stratifications, and not reported these analyses or results and the researchers mention that p-values were not adjusted for multiple testing.

While this research may not present a clear-cut case of p-hacking, it does suggest that trying out different combinations of RDFs is warranted in instances where needs are urgent. At the time, there were no available treatments that reduced mortality in patients with severe COVID-19 and already scarce healthcare resources were becoming increasingly under threat. Regarding how it decided on its guidelines, the WHO states:

“The panel ranked the outcomes and attributed a high value to even a very small reduction in mortality. In addition, the panel also placed a high value on even a small reduction in the need for mechanical ventilation, which places a large physical burden on patients and an emotional burden on patients and families. A second reason the panel placed a high value on a small reduction in mechanical ventilation concerns health resource issues: the availability of mechanical ventilation stands out as an important vulnerability during the COVID-19 pandemic.” (WHO Living Guidance 2020: 9)

Moreover, there was less concern for potential harms caused by the drug given that the relative safety of the drug had previously been established, and the treatment was readily available and reasonably priced. That said, the WHO categorized their recommendation for dexamethasone as “strong” but indicated that it was “based on moderate certainty evidence”. This seems indicative of tolerance for a somewhat looser evidence base than typically desired for their guidelines.

While the AIR does not necessarily defeat the strict conditions of the prevalent position, it does, to a certain degree, vindicate the argument that choosing to p-hack can, under certain circumstances, be warranted. That is not to say that p-hacking is valuable for establishing the truth of hypotheses, but rather for discovery, particularly when the stakes are high. Non-epistemic judgements of inductive risk can warrant decisions to p-hack. Furthermore, these considerations imply that the prevalent position may be too strict a view of how researchers should conduct research, particularly when non-epistemic values of wrongly rejecting a hypothesis are severe.

3.5 Conclusion

The analyses and arguments I have offered have potential implications for future discussions of p-hacking. For one, it seems plausible that different interpretations of p-hacking will have some influence on the techniques for its detection. Indeed, those techniques which claim to be detecting p-hacking, may in fact be detecting a range of questionable practices. Another area of philosophical interest involves the role of funding in encouraging or discouraging p-hacking and related practices, particularly when it comes to the interaction between epistemic and financial resources. Finally, during the course of the COVID-19 pandemic in particular, much has been said about increased amounts of “bad” science (see Scheirer 2020; New Scientist 2020). Once it has passed, it would be interesting to investigate the extent to which p-hacking contributed to this dissemination of misinformation. To be sure, tackling these issues relies on knowing how to characterize p-hacking, how severe its consequences can be, and when it is warranted.

I set out to achieve three things in this Chapter. First, I unpacked the different features of p-hacking to provide a much-needed explicit definition of the practice and clarify its importance for efficacy claims and medical inference. P-hacking is characterized as the intentional or unintentional abuse of researcher degrees of freedom, which may lead to exaggerated results regarding a particular hypothesis of interest.

Second, I articulate the prevalent position on p-hacking and, in doing so, also illustrate just how serious the consequences of p-hacking can be. While p-hacking is widely condemned on epistemic grounds, it is also important to consider its practical consequences. My expected utility approach captures both concerns. In this regard, I appeal to influential work on false-positive report rates and p-hacking to argue that just four decisions to exploit researcher degrees of freedom entails an extremely high probability that a given positive result is false. Applying tools from expected utility theory illustrates how the enormity of the epistemic effects can have significant practical consequences. Moreover, these tools clarify precise conditions under which p-hacking is practically harmful: Just in case p-hacking increases the probability of obtained positive results being false, and acquiring true-positive results will have more beneficial outcomes than acquiring false-positive ones, then we should choose not to p-hack.

Finally, I argue that analytic choices regarding researcher degrees of freedom can sometimes be appropriate. I analyze whether considerations of inductive risk can be used to warrant p-hacking. While it is unavoidable that p-hacking increases false-positive report rates, non-epistemic judgements of urgency can warrant decisions to p-hack.

Chapter 4

Curb Your Effectiveness: Correcting for Meta-Biases in Therapeutic Prediction

4.1 Introduction

It is commonly assumed that results from medical research can be used to determine the general capacity of medical interventions to bring about their intended effects. This is done by inferring from measurements of treatment efficacy to estimations of treatment effectiveness. In this Chapter, I argue that such estimations are prone to failure because they do not account for the effects of a particularly nefarious set of biases on measures of medical effectiveness: *Meta-biases*. To help remedy this, I offer a novel model for correcting the results of clinical studies based on evidence about these meta-biases. I refer to this framework as the *bias dynamics model*.

Clinical research is performed to measure the efficacy of medical treatments. When conducting a clinical trial, researchers quantify the effect of an intervention as an effect size using some outcome measure.²⁰ This is a measure of the net difference that exposure to an intervention makes to a particular outcome in a population. Effect sizes are calculated by performing statistical tests on data gathered in clinical trials, such as cohort studies or randomized control trials (RCTs). If a trial meets certain standards, the measured effect size is thought to accurately quantify the relationship between exposure to the treatment and the outcome for the study population in the trial. In other words, the study will exhibit a high degree of internal validity. The results from clinical trials are subsequently used to predict the effectiveness of medical interventions. In other words, measures of medical efficacy are used to estimate the general capacity of medical interventions to bring about their desired effects in populations outside of trial settings.

One major challenge facing clinical trials is that, even if they are internally valid, they may nevertheless fail to be externally valid—the conclusions drawn about the study population may not be true of the target population. Put another way, an effect measure can be true for a sample population but not true when it is generalized to a target population. Establishing the external validity of clinical trials is a crucial concern in medicine, having received a lot of attention from medical researchers, methodologists, and philosophers alike. One proposed method for establishing external validity is to develop a reliable approach to *extrapolation* by outlining the principles under which we would be justified in generalizing an effect size from a sample population to a target population.

²⁰ See Appendix 2 for a breakdown and explanation of commonly used outcome measures in medicine.

There are currently two dominant approaches to extrapolation proposed in the medical and philosophical literature. The first is what Stegenga (2018) refers to as simple extrapolation unless (SEU). According to SEU, generalizing a result from a clinical trial is admissible if there is no good reason to believe that the target population differs from the population in the trial in some relevant way (Post et al. 2013; Walsh et al. 2015; also see Chapter 2). Proponents of this approach argue that, usually, no reason will be found and thus we can extrapolate the findings from clinical trials without risk (Guyatt et al. 2008). The second approach is to explicitly establish the sufficient similarity of the sample and target populations before extrapolating research results. Generalizing a result from a clinical trial is admissible, on this approach, if it has been established that the target population is shares salient features with the sample population from the trial. This strategy has received more philosophical attention in recent years, with several different approaches to what counts as relevant for establishing such similarity and how best to do so (Steel 2008; Cartwright 2012; Broadbent 2013; Fuller and Flores 2015; Parkkinen et al. 2018).

While these two approaches to extrapolation provide possible solutions to the problem of external validity, they rely heavily on having accurate and reliable results from clinical trials in the first place. Typically, discussions about the reliability of clinical trial results center on biases that may manifest in the process of the trial itself and which threaten the internal validity of medical research. Such biases include confounding, confirmation bias, reporting bias, and various forms of data-dredging. However, a problem that is often overlooked in discussions about the application of results from clinical trials is another form of biases that affect medical research at the meta-level, and thus only indirectly affect the reliability of research results. I refer to these biases as meta-biases. Widespread meta-biases present a serious challenge to the reliability of medical evidence, particularly in cases where that evidence is used to estimate the effectiveness of medical interventions. Negative analyses and studies are often relegated to the file drawer, and industry sponsored research tends to promote commercial interests. Some researchers argue that the ubiquity of research biases contributes to a large proportion of published scientific conclusions being false (Ioannidis 2005; Horton 2015). More recently, others have suggested that these such biases, including meta-biases, can, and typically do, lead to systematically exaggerated claims of medical effectiveness (Stegenga 2018; Fuller 2021).

My first aim in this Chapter is to expand on this line of reasoning, by arguing that, despite a growing body of evidence on their impacts, we often fail to account for the effects of meta-biases in medical research. That is, despite evidence that estimations of medical effectiveness based on the results of clinical trials are often overstated due to pervasive meta-biases, we do not have procedures to correct for their effects. My second aim is to offer a strategy to help remedy this problem. I outline a framework for correcting measures of efficacy in line with empirical evidence about the prevalence and effects of meta-biases prior to inferring claims about effectiveness. The model I propose is loosely based on the methods used in the field of dynamics, a branch of classical mechanics, whereby calculations of the forces impinging of the motions of objects are adjusted to account for friction.

Just as such calculations employ empirical *friction coefficients* to modulate estimates of the forces acting on objects, I propose the use of empirical *bias coefficients* to adjust estimates of medical effectiveness. Such bias coefficients, I argue, should be regularly updated to reflect the best current evidence about meta-biases—they thus change over time. Because of its relation to dynamics calculations, and because I propose that bias coefficients be amended frequently based on empirical findings, I refer to my proposed framework as the *bias dynamics model* for estimating the effectiveness of medical interventions.

I begin, in Section 4.2, by providing a definition of meta-biases. Here, I draw a key distinction between what I refer to as *methodological biases* and the higher-level concept of meta-biases (Section 4.2.1). Methodological biases are those biases that are directly linked to the methodological features of clinical research. Meta-biases, on the other hand, are connected to meta-level properties of a scientific discipline. This distinction helps demonstrate that disciplinary guidance often focuses on methodological biases and often overlooks the effects of meta-biases. In Section 4.2.2, I argue in line with other epidemiological and philosophical work, that inferences of medical effectiveness are often exaggerated due to the impacts of these types of biases. I appeal to two prominent meta-biases, *publication bias* and *sponsorship bias*, to illustrate the point that the problem of meta-biases should lead us to attenuate our estimations of the effectiveness of medical interventions. Meta-biases lead to systemic exaggerations of research results, which are then used to make inferences about the effectiveness of medical interventions. Thus, such estimations are prone to unreliability, regardless of whether the principles of simple extrapolation unless or extrapolation by sufficient similarity are satisfied. In Section 4.3, I outline the bias dynamics model, a strategy that corrects for meta-biases when inferring effectiveness. I argue that estimations of effectiveness can be attenuated using what I refer to as *bias coefficients*, which represent the effects of meta-biases on specific outcomes in particular subdomains of medical research. I then briefly outline a provisional research program for calculating bias coefficients in Section 4.4. In line with novel contemporary work in meta-science, I propose the use of simulation studies to help model the effects of meta-biases on measurements of treatment efficacy based on existing research on the prevalence of such biases. Following this, I conclude.

4.2 What are Meta-Biases?

Medical research is plagued by various forms of bias. The Cochrane Group defines a bias as “a systematic error, or deviation from the truth, in results or inference” (Cochrane Handbook 8.2.1) Biases are typically thought to occur due to some property of either the design or implementation of a research method or the interpretation of the data gathered through experimentation. It is worth asking whether this conception of bias fully captures the full range of ways in which medical research can be systemically distorted. Below, I draw a useful distinction between the standard characterization of biases, which focuses on methodological aspects of clinical research, and biases that occur at the meta-level of scientific disciplines.

4.2.1 *Methodological Bias and Meta-bias*

Standard examples of bias include confounding, confirmation bias, the Hawthorne effect, and reporting bias. Biases like this are relatively well-understood. They are directly linked to methodological features of clinical research. As such we have a good idea of how they manifest and have invested a great deal into developing strategies to mitigate and prevent them. Take confounding, the underdetermination of the association between a treatment and an outcome due to an imbalance of some factor in the experimental and control groups of a study. Some strategies for lowering the chance of confounding in clinical research are randomization, matching, stratification, and multivariate analysis. Another oft-cited example of bias is confirmation bias, a catchall term for various ways in which one tends to give more weight to evidence that supports their hypothesis than to evidence against that hypothesis. Good experimental blinding and having rigorous procedures for finding, reporting, and correctly weighting all relevant evidence are widely endorsed practices for preventing confirmation bias. Because of their direct relation to the processes within clinical trial methodology, these types of bias may be referred to as *methodological biases*.

Perhaps the most damaging form of methodological bias is *reporting bias*. This bias occurs when results are distorted due to the selective disclosure of the analyses performed and results obtained in a clinical trial. Practices that lead to reporting bias include withholding unfavorable or nonsignificant results, publishing only a subset of the analyzed data, reporting secondary outcomes as primary outcomes when the original primary outcomes yielded nonsignificant results,²¹ and adding entirely new outcomes to the published study. While difficult to detect, there is evidence that practices that constitute reporting bias are widespread in clinical research. Between October 2015 and January 2016, The COMPare Trials Project, run by the Centre for Evidence Based Medicine Outcome Monitoring Project, found that just nine of 67 published clinical trials in leading medical journals were reported without error (Goldacre et al. 2016). The implication here is that reporting bias, whether intentional or not, is rampant in medical research. Elsewhere, a six-month study of top medical journals, including *The Lancet*, *British Medical Journal*, the *Journal of American Medical Association*, *Annals of Internal Medicine*, and *New England Journal of Medicine*, found that 18% of RCTs had inconsistencies in their reporting of primary outcomes when compared with corresponding trial registry entries (Fleming et al. 2015). The researchers also found that 64% of trials had discrepancies relating to non-primary outcomes, including the omission of outcomes listed in registry entries and addition new.

Because of the scrutiny reporting bias has received, the mechanisms responsible for it are relatively well-understood and strategies to prevent it have been developed with some success.

²¹ A primary outcome is a variable that is the most relevant to answer a research question; it is the main measure in a study. A secondary outcome is an auxiliary variable which is measured to help interpret the primary outcome. Sometimes there are more than one primary and secondary outcome. It may be the case that an intended primary outcome does not yield positive results, but a secondary outcome does. Without a pre-registered study protocol, it is difficult to determine if the intended primary outcome has been abandoned and replaced by a secondary outcome.

Requiring that studies be recorded on official trial registries is thought to greatly mitigate outcome reporting bias. Some studies show that there was a reduction in some of the mechanisms that lead to the bias, such as questionable design and analytic practices, after trial registration was introduced at the turn of the millennium (Kaplan and Irvin 2015). However, there is evidence that reporting errors have, at least in some fields, remained consistently high (Nuijten et al. 2016; Howard et al. 2017), and a quick comparison of research on the rate of selective reporting in Cochrane reviews shows that the risk of bias has increased (cf. Kirkham et al. 2010 and Shah et al. 2020). Therefore, preregistration, at least in its current form, is evidently insufficient for preventing outcome reporting bias. However, new, and updated efforts to decrease outcome reporting bias are being pursued.

Methodological biases like those described above are emphasized in influential EBM guidelines. For example, the Cochrane Group offers a tool to assess the ‘risk of bias’ for studies intended for inclusion in systematic reviews (Cochrane Handbook 8.5.a). The tool lists six broad categories of bias: Selection bias, performance bias, detection bias, attrition bias, reporting bias, and a final category for other bias. Researchers are advised to assess the risk of a particular bias by looking at specific methodological aspects of the clinical trial in question. For instance, to assess the risk of detection bias, one should assess the extent to which the analysts in the study were blinded in their assessments of each outcome. Confounding, confirmation bias, and detection bias (and other similar biases) fit well under the standard characterization of bias in medicine. They are indeed linked to problematic procedures in data collection, analysis, and interpretation.

Another set of biases, however, cannot be so explicitly connected to features of the experimental process. Rather, these biases are linked to meta-level properties of science, such as entrenched values of the relevant scientific community or inveterate norms of its research system. Because of their connection to meta-level properties that are external to experimental methodology, I will refer to these biases as *meta-biases*. Of course, meta-biases can influence how research is conducted—they may indeed precipitate or manifest via questionable practices such as p-hacking (see Chapter 3) or selective reporting—yet their effects cannot be seen by looking at single studies, as one would the effects of methodological biases. The distortions in results due to meta-biases are only observed when comparing sets of studies with particular meta-level properties.

An exemplar of meta-bias is *publication bias*. After completing a clinical trial and analyzing the data, researcher will decide whether to publish the results. The publication of clinical trial results is skewed by decisions not to publish results: It is common for negative outcomes, results that indicate no causal relationship between a treatment and an outcome, to go unpublished. There is strong evidence for this. Murad et al. (2018), for instance, have found that studies with positive results are 3.90 times more likely to be published than those with negative results. This leads to a general weighting toward positive results in published medical research. Here, the distortion may be observed by comparing the set of all published studies to the set of all completed studies.

Another example of meta-bias involves observed differences between the findings of industry funded studies and those of non-industry funded studies. The majority of clinical trials are funded

and conducted by private organizations that have vested interests in the outcomes of those trials. There is evidence that industry funded studies tend to generate a higher ratio of results that promote the interests of the funding organization when compared to non-industry funded studies.²² A recent Cochrane systematic review concluded that industry sponsored trials were 1.27 times more likely to report beneficial outcomes, 1.34 times more likely to show less evidence of harms, and 1.37 times more likely to present more favorable overall conclusions when compared with non-industry funded studies (Lundh et al. 2017). The tendency of clinical trials to generate results that promote a sponsor's interests know as *sponsorship bias*.²³

Understanding that sponsorship bias is a meta-bias is somewhat more complicated than the more straightforward case of publication bias. This is because it is perhaps common for methodological biases, such as reporting bias, to occur in efforts to generate findings that benefit a trial sponsor. Nevertheless, there is clear evidence that the meta-level properties of studies, namely 'industry funded' and 'non-industry funded', are associated with a broad systemic distortion towards positive results that favor trial sponsors' interests. Indeed, industry sponsorship is not a necessary feature of trials where related methodological biases occur. Nor is it necessarily the case that researchers in industry sponsored trials with positive findings have committed methodological biases. It is simply that industry sponsored trials tend to generate positive findings in favor of sponsors' interests. And since most trials are sponsored by industry, and most industry-funded studies report positive findings in favor of their sponsor, there is good reason to believe that industry sponsorship leads to a general positive skew in research results.

Still, it may be argued that, on its own, evidence that industry sponsored trials tend to report beneficial outcomes when compared to non-industry funded trials is insufficient for establishing the existence of the funding effect. After all, pharmaceutical companies will often halt research programs that run the risk of failure before they reach clinical trials. And non-sponsored research is often slow of the mark in this regard. This provides an ostensible explanation for the higher rates of positive results from sponsored trials. However, there is evidence that sponsored head-to-head trials tend to generate results that favor the sponsor's drug over a competitor when compared to non-sponsored head-to-head trials (Ridker and Torres 2006; Flacco et al. 2015). In other words, industry funded trials are more likely to conclude that their drug is superior or, at least, not inferior to competitor drugs than non-industry funded trials. More precisely, industry funded trials tend to find that the sponsor's drug is better than other competitor drugs, while non-industry funded tend to find that those drugs are not superior competitors. This failure to reach similar conclusions regarding particular treatments further supports the existence of sponsorship bias in medical research.

A major issue facing the problem of meta-biases is that they have not been explored to the same extent as methodological biases. As a result, strategies to help prevent them are nascent and somewhat limited. The Cochrane risk of bias tool gestures toward this concern in its 'other bias'

²² See, for example, Bekelman (2003) Lexchin et al. (2003); Sismondo (2008a); Lundh et al. (2012).

²³ Sponsorship bias is also known as industry funding bias, funding bias, and the funding effect.

category, recommending that systematic reviewers simply “State any important concerns about bias not addressed in the other domains in the tool” (Cochrane Handbook 8.5.a.). This leaves much open to interpretation and risks allowing further subjectivity into the process of evidence amalgamation (see Stegenga 2011).

Failing to commit resources to researching meta-biases risks hindering efforts to limit their effects. In order to prevent publication bias, for instance, some journals have made it policy to solicit and publish research with negative results. However, there are other factors that lead to publication bias besides the historical tendency of journals favoring positive results. For instance, positive results are more likely to be cited and having highly cited results increases the one’s change of funding and career promotion. Other initiatives to mitigate publication bias include governmental and organizational policy requiring that the results of certain studies be reported on clinical trial registries within a year of completion be passed.²⁴ Yet, such laws are often undercut by non-compliance on behalf of researchers and regulatory loopholes (Goldacre et al. 2018).

Another more recent strategy, the use of funnel plots, is directed at detecting publication bias. This formal method, as Holman (2019: 12-13) notes, does not require access to unpublished studies and is routinely used in medical research. While useful for assessing the risk of publication bias in meta-analytic research, funnel plots do not provide a way to correct for its effects. Researchers who utilize funnel plot analysis and find that there is a risk of bias typically mention this risk as a caveat in the discussion sections of their studies. Thus, while helpful, funnel plots do not directly tackle the challenge of publication bias, but rather further reveal the extent of the problem. Overall, publication bias is a multifaceted problem and preventing it involves immense coordinated efforts.

The same can be said of sponsorship bias. Doucet and Sismondo (2008) describe five interconnected sources of sponsorship bias, including elements of trial design, the practice of performing multiple trials with predictable outcomes, outright fraud, the manipulative use of rhetoric in published articles, and overlap with publication bias. Proposed solutions to sponsorship bias include financial disclosures, standardized reporting, and trial registration. However, these policies fail to cover all the contributing factors (Doucet and Sismondo 2008). Financial disclosure has been common practice for many years, yet industry funded trials continue to regularly favor sponsors’ interests. Standardized reporting cannot fully address trial design concerns. And while it may make fraud more difficult, developing strict reporting guidelines will not fully deter those whose goal is to manipulate data through outright fraud. Furthermore, the complexities of rhetoric in articles are hardly solved by the introduction of reporting standards. Such standardization may contribute to what Steel (2018) refers to as *inferential asymmetries* in the interpretation of clinical results, whereby some stakeholders are less able to infer true conclusions than others due to the way in which research results are reported. Trial registries may help solve sponsorship bias as it relates to publication bias, but one could follow registration procedures correctly and still introduce industry

²⁴ See Food and Drug Administration Amendments Act of 2007, European Medicines Agency (2014), and National Institutes of Health, Department of Health and Human Services (2016).

favoring design features, perform multiple studies, directly manipulate data, and present findings in a way that promotes the sponsor's interests.

Ultimately, sponsorship bias cannot be so easily reduced to issues of methodology. It is doubtful that the five contributing factors outlined by Doucet and Sismondo (2008) are all there is to sponsorship bias. The funding effect can manifest via multiple first order methodological biases or through outright fraud. It is, in this sense, multiply realizable—in individual trials, different constellations of biases may be responsible for generating results that favor sponsor interests. However, it is often difficult to know whether methodological biases have occurred in a trial, and when we do know, it is somewhat challenging to explicitly connect these to a desire or tendency to favor a funding organization's interests. Moreover, existing quality assessment tools (QATs) for evaluating evidence, which include assessments of the risk of bias for particular studies, often have poor inter-rater and inter-tool reliability leading to uncertainty regarding how sources of evidence should be weighed against each other (Stegenga 2015c; 2018). In other words, there are often epistemic gaps regarding first order methodological biases connected to the sponsorship bias and their effects on the quality of evidence. Here, the higher order concept of sponsorship bias can be useful since we can have clearer evidence to show how industry funding is correlated with favorable results.

Sponsorship bias can, in this sense, be thought of as a more distal bias responsible for the systemic distortions in clinical research in comparison to the more proximal methodological biases that it precipitates or from which it manifests. Indeed, it seems like the funding effect is a problem of bad barrels rather than bad apples—when it comes to results skewed in favor of industry, it is the entrenched seating of private sponsors as the majority curators of clinical research, not simply pervasive methodological biases, that is responsible for the general disparity in results. That is what makes sponsorship bias a meta-bias. Thus, preventing sponsorship bias involves the mammoth task of completely overhauling clinical research, with some suggesting that private sponsors be excluded from research (Angell 2005; Sismondo 2008b). However, even if this were possible, it would require an enormous amount of time and resources, all while the effects of sponsorship bias persist.

This discussion should make three things clear. First, it is often helpful to use the higher order concept of meta-biases. While it might be the case that a particular meta-bias is constituted by methodological biases, we may not have epistemic access to those first order biases. Second, the problem of meta-biases is difficult to solve because such biases are typically linked to structural-level properties of a scientific discipline. Current strategies for the prevention of meta-biases require more revision and evaluation, and even when successful, the results take time. Despite this, research on meta-biases can be utilized in our inferential practices. In Section 4.4, I offer an alternative to prevention, one that corrects for the effects of meta-biases on clinical research results that are used in extrapolation. Prior to this, it is worthwhile explaining the effects of meta-biases on extrapolation in medicine. Doing so makes clear that there is the need for methods directed at correcting for the effects of meta-biases.

4.2.2 *How Meta-Biases Affect Estimations of Medical Effectiveness*

The distortions in results caused by pervasive meta-biases are a serious problem for medical research. For example, a growing body of evidence shows that publication bias leads to a systemic overestimation of effect sizes generated through meta-analyses of clinical research (Hopewell et al. 2009; Kicinski 2014; Murad et al. 2018). Because there is a higher proportion of positive findings than negative findings in the published literature, the pooled results of the published studies will show greater effectiveness than if all studies were included. Likewise, we can infer from evidence about sponsorship bias that it generally skews results such that medical effectiveness is overestimated. If most clinical trials are industry funded and if industry funded trials tend to generate industry favoring results when compared to non-industry funded trials, then this gives us good reason to believe that (a) any given industry funded trial is more likely than not biased, and (b) amalgamations of evidence from all relevant published trials will bake sponsorship bias into their results. Because of this, there are principled reasons to lower our confidence in the results of a particular industry funded RCT and meta-analyses that include industry funded trials.

The impact of meta-biases is not restricted to evidence amalgamation techniques but extends to extrapolation. Here, the problem is that extrapolation involves estimating the effect of an intervention in some target population based on what we take to be the true effect in the sample population. However, because of the amplifying impact of meta-biases, the effect in the sample population of a meta-analysis could in fact be less than that which was measured. Notably, the meta-analysis could include risk of bias assessments, and include funnel plots for detecting publication bias, and still the results may be used to extrapolate. If there is no correction to the results based on evidence of meta-biases, the extrapolation will fail for reasons antecedent to there being relevant differences between populations. Stegenga makes clear why publication bias, in particular, is a problem for simple extrapolation unless:

Even if there are in fact no substantial differences between the experimental subjects and target patients—and thus the overriding clause of SEU were not satisfied, and so extrapolation could be warranted—the results of published trials from which one is extrapolating could be entirely misleading, because the published trials may represent only a fraction of the trials that have been performed (2018: 127)

Stegenga's argument has an ontological character. He notes that even if there were *in fact* no relevant differences between the study and target populations, publication bias would be a problem. The argument can be couched in epistemic terms too: Even if we have *no good reason to believe* that the study and target populations were different in relevant ways, and thus according to SEU extrapolation is justified, publication bias would be a problem. Either way, because positive results are generally favored for publication, the results from the set of published trials will report a higher degree of efficacy than if unpublished trials were included, thus extrapolating effect sizes derived in meta-analyses of those studies would result in inflated claims of effectiveness in the target.

This argument can be extended to other meta-biases and extrapolation by sufficient similarity too. Even if (there is good evidence establishing that) the individuals in the target population are sufficiently similar in all relevant respects to that of the pooled published studies, and thus extrapolation is warranted by ESS, the favoring of positive results over negative results for publication will lead to exaggerated estimates of effectiveness in the target population. There will be similar consequences given sponsorship bias. The results of industry funded studies, which make up the majority of clinical trials, typically report outcomes that favor the sponsor's treatment. Thus, the set of trial results being extrapolated could be positively distorted due to the tendency of published studies to favor industry treatments, even if either SEU's or ESS's principle of extrapolation is satisfied. In other words, the problem and other meta-bias transcends concerns about the internal and external validity of clinical trials. A study may score high on internal validity by satisfying typical methodological criteria, such as whether it randomized its participants and whether it was double-blinded, and may also satisfy the principles of extrapolation aimed at external validity set out by SEU or ESS, and yet the results may deviate from the true effect of the intervention due to meta-biases.

Of course, this is not to say that the problem of meta-biases is a direct problem for extrapolation in medicine, but rather that if we do not have methods to correct for their effects, then extrapolations will likely fail. Moreover, without meta-bias corrections, it would be more difficult to assess whether such failures of extrapolation are due to faulty measurements of efficacy or due to there being relevant differences between populations. This provides more reason for medicine to have ways to correct for meta-bias effects.

I will not belabor the point further. Meta-biases present a major problem for estimating medical effectiveness, and, more indirectly, for extrapolation, because they lead to exaggerated measures of treatment efficacy. In the absence of effective strategies for preventing meta-biases, there should be a way to attenuate measures of efficacy based on what we know about the influence of meta-biases. In the next section, I describe a model for how this can be done and outline how research on the incidence and effects of meta-biases can be utilized to modulate estimations of medical effectiveness.

4.3 Correcting for Meta-Biases: The Bias Dynamics Model

In this section, I offer a model to correct for the effect of meta-biases on estimations of medical effectiveness. The model I propose is loosely based on friction force dynamics calculations in physics. In these calculations, empirical *friction coefficients* are used to estimate the effect of friction between two surfaces in calculations of the motions of objects. I propose a similar approach for predicting medical effectiveness, whereby empirical *bias coefficients* can be used to estimate the effect that meta-biases have on measures of medical efficacy. I refer to my model for correcting for meta-bias as the *bias dynamics model*.²⁵

²⁵ The bias dynamics model draws inspiration from Stegenga's proposals in Appendix 5 of his book, *Medical Nihilism* (2018).

Assume a meta-analysis is conducted to measure the effect of some treatment on some dichotomous outcome. After the data is collected and the analyses are completed, the researchers calculate an effect size measured as an absolute risk reduction (ARR). In keeping with established terminology, the ARR counts as evidence that supports what is referred to as a claim about *efficacy*—a proposition about the capacity of an intervention to bring about its intended outcome in a group of individuals in a particular population, such as one in a study setting (see Cartwright 2009). I will use ARR_S to refer to an effect size, measured as an absolute risk reduction, that stands as evidence for a claim about efficacy, one calculated for a given study population.

Of course, when conducting such a study, we are not just interested in whether the intervention is effective for the study population, but also whether it will be effective more generally. In other words, we are interested in what is referred to as a claim about *effectiveness*—a proposition about the capacity of an intervention to bring about its intended outcome in populations other than the one in the study, some target population. Since we usually do not have direct evidence about the extent to which an intervention will work in a target population, we estimate effectiveness using the empirical results from clinical studies. The effect size from a study, in this case ARR_S , is used to estimate the general capacity of the treatment in the target population. I will use ARR_T to refer to the estimated effect size for the target population, the claim about the effectiveness of the intervention.²⁶ This target effect size represents an expectation about how effective the intervention will be once meta-biases have been considered. We can treat ARR_T as an *expected frequency*, representing the predicted rate of at-risk individuals who would benefit from the intervention in question.

The aim here is to take what we believe to be a true measure of a treatment's effect in the study and infer the general capacity of the intervention that is as close to the truth for a target population. Many, including those in the EBM movement, assume that, barring threats to internal validity and relevant differences between sample and target populations, we can straightforwardly apply such results. Thus:

$$ARR_T = ARR_S$$

However, as I and others have argued, pervasive meta-biases should lead us to attenuate clinical trial results—there is good reason, besides differences between populations and traditional failures of internal validity (i.e., methodological biases), to expect that the effectiveness represented by ARR_T is lower than the efficacy represented by ARR_S .

²⁶ While there are concerns about the generalizability of effect sizes measured as absolute risk reductions (see Glasziou and Irwig 1995), I do not have the space to deal with them here. It is worth noting that Glasziou and Irwig recommend using relative risk over ARR as a measure of effectiveness because it is more mathematically reasonable. Fuller (2021) provides an argument against the use of relative risk, in the same vein as that used against ARR. And Stegenga (2018) recommends the use of absolute outcomes measures, in particular ARR. Of course, one's choice of effect size is relevant to the reliability of extrapolation, however that is not of immediate concern here since the bias dynamics model can be modulated to work with different outcome measures.

Lowering the expected effect size can be done by introducing what I refer to as a *bias coefficient*, δ , where $0 \leq \delta \leq 1$. Ideally, the bias coefficient represents the effect of all meta-biases on effectiveness claims. By multiplying the effect size obtained in the trial by the bias coefficient, we lower the expected effectiveness of the treatment in question:

$$ARR_T = \delta ARR_S$$

Since the magnitude of the bias coefficient is between 0 and 1, it will always attenuate the value of ARR_T . The closer δ is to 0, the more meta-biases affect ARR_T , and likewise, the closer δ is to 1, the less meta-biases affect ARR_T . It is theoretically possible that the effects of meta-biases are sufficient ($\delta = 0$) for us to expect that the intervention will have no effect in the target ($ARR_T = 0$). On the other hand, meta-biases could, in principle, have no effect ($\delta = 1$), and thus we could conclude that the estimated effectiveness in the target will be equal to the measured efficacy in the trial ($ARR_T = ARR_S$).²⁷ However, given the prevalence of meta-biases, these scenarios are unlikely in practice.

The bias dynamics model provides a way for researchers to correct measures of effectiveness by taking the effects of meta-biases into account. As mentioned, the bias coefficient ideally represents the effect of all meta-biases on research results, yet, in practice, it is very unlikely that we would have access to such knowledge. This does not, however, preclude the use of bias coefficients. We can determine bias coefficients based on what is known about meta-biases. This is the next step to calls for the evidence we gather about the medical research system and its practices—typically the purview of the burgeoning field of *meta-research*—to be used for adjusting confidence in clinical findings (Ioannidis 2008; Fuller 2018a; Stegenga 2018). Furthermore, bias coefficients can fill the gap left in recognized guidelines on biases by helping researchers account for the effects of meta-biases, which are not explicitly listed in quality assessment tools (QATs). That is to say, once a QAT, such as Cochrane’s risk of bias tool, has been used to rate single studies for bias, the bias coefficient can be used to determine an overall rating for a given research program based on evidence about meta-biases. In line with this, an organization like Cochrane might, for instance, publish domain specific bias coefficients in their guidelines. Furthermore, at least in the case of publication bias, bias coefficients can be used in conjunction with funnel plot analyses. Here, researchers may detect publication bias and have more justification for the use of a correcting bias coefficient based on their findings.

Bias coefficients should be categorized by research area and outcome. Meta-biases occur at various rates in different research programs—the prevalence of outcome reporting bias in cancer research differs to that in research on statins (cf. Vera-Badillo et al. 2016 and Rezende et al. 2018). Likewise, different outcomes within each research program will have different rates of meta-biases—the rate of publication bias in research on the association between statins and heart attack may differ

²⁷ Assuming there are no other more direct threats to internal validity and that our accepted principles of extrapolation are satisfied.

from that research on the association between statins and stroke. Naturally, because of this, the values of bias coefficients for given outcomes in particular fields will differ.

Unlike their analog in physics, bias coefficients should be dynamic. That is to say, the values of bias coefficients should be updated over time with the observation of new evidence about meta-biases. Such an approach is in keeping with the central tenets of EBM and its related organizations, which aim to update their guidelines based on the best available evidence. For example, the Cochrane Group has already gestured toward this sort of approach with updates to its research into the effects of industry funding on research outcomes. In 2012, evidence showed that industry funded studies were 1.32 times more likely to report results that favored the sponsor's interests (Lundh et al. 2012). This study was updated to reflect the 2017 evidence-base, finding that industry funded studies were 1.27 times more likely to report favorable results (Lundh et al. 2017). Regular updates to bias coefficients in line with research like that cited here will provide researchers with a means to provide more accurate estimations of medical effectiveness.

To briefly demonstrate the bias dynamics model, consider a study by Chou et al. (2016) in which several meta-analyses of evidence for the effects of statin therapy on various outcomes were conducted. In the study, one meta-analysis of 12 clinical trials found that statin therapy is associated with a 0.81% decrease in heart attacks over six years. That is, the absolute risk reduction of heart attack when taking statins for six years for the pooled participants from these trials is 0.81%, a beneficial outcome. Thus, in the statin study $ARR_S = 0.81\%$. We want to know if using statins will decrease the risk of heart attack in more general populations of individuals with high cholesterol, a risk factor for heart attack; in other words, we want to estimate the general effectiveness of statins in decreasing the risk of heart attack. Most would assume that $ARR_T = 0.81\%$. We can expect that around 0.81% of individuals in other populations of individuals with high cholesterol will benefit from statin therapy. As argued though, evidence about meta-biases in statin research should be considered.

While there is little indication about the extent of publication bias in statin research, others have found evidence of serious industry funding bias. For instance, when it comes to RCTs comparing competing statins for cholesterol reduction, Bero et al. (2007) found that the odds of statistically significant results in favor of statin therapy in drug-drug comparisons were 16 times greater for industry funded trials than for non-industry funded trials. While these findings do not indicate exactly how much meta-biases affect statin research regarding one's risk of heart attack, they do warrant assuming a relatively low bias coefficient (indicating a large effect from meta-biases). In line with this, assume that research of the kind I outline in Section 4.4 below is performed to determine the bias coefficient for statin research on cholesterol reduction. Now assume that this results in an approved bias coefficient for research into the effect of statin therapy on one's risk of heart attack of, say, $\delta = 0.45$. This would attenuate the expected frequency of individuals who would benefit from statin therapy, $ARR_T = 0.37\%$.

One might wonder why I do not propose applying my framework at the level of methodological biases. This is, after all, a compelling route to take. Having bias coefficients that represent the effects of methodological biases may, in principle, generate more accurate claims of effectiveness. Yet, there is at least one practical reason to prefer employing the bias dynamics model at the level of meta-biases. As previously discussed in Section 4.3.1, there are often epistemic gaps relating to first order methodological biases. They are difficult to detect and thus it is challenging to assess their prevalence and effects, which is key to generating bias coefficients. This, added to commonly poor inter-rater and inter-tool reliability of evidentiary quality assessments based on judgements about methodological biases (Stegenga 2015; 2018) makes achievability of the accuracy hoped for by applying the model at the first order level unlikely. Of course, employing the bias dynamics model at the level of meta-biases also provides only approximations of effectiveness. But the type of evidence used is less susceptible to the kinds of epistemic gappiness we currently find at the level of methodological biases. Furthermore, many meta-biases manifest via methodological biases, and thus correcting for meta-biases indirectly corrects for at least some methodological biases.

The bias dynamics model provides a straightforward strategy for researchers correct for the effects of meta-biases. The bias coefficient functions as a regularly updated attenuating variable representing the known effects of meta-biases. But how should the value of the bias coefficient be determined? In the next section I outline several plausible routes to determining bias coefficients.

4.4 A Provisional Proposal for Determining Bias Coefficients

Determining bias coefficients involves conducting meta-research on the incidence and effects of meta-biases in various subdomains of medicine. Studies like those of Murad et al. (2018) on publication bias and Lundh et al. (2017) on sponsorship bias are important for ascertaining how prevalent these biases are, yet these studies do not provide estimates of the differences in measured effects due to meta-biases. Perhaps the best way to determine the influence of meta-biases is to perform clinical trial type studies on existing clinical trials. RCTs are lauded by those in the EBM movement as the gold-standard of evidence (GRADE Working Group 2013). However, conducting RCTs on the effects of meta-bias is somewhat of a non-starter since this would mean introducing a meta-bias (the intervention) into a population of trials randomized to a treatment group (who receive the meta-bias) and a control group (who do not to receive a meta-bias) to understand the effect of the meta-bias. It would be impossible to implement the mechanics of such an RCT in a real-world situation since we cannot idealize the situation in the right ways. We cannot run multiple trials for inclusion in our trial, some of which we introduce a meta-bias to, and others we do not. And running such an RCT on real world trials would be both epistemically and ethically questionable; given that trials are themselves meant to be free from meta-bias, we cannot go introducing some for the purpose of meta-research.

A reasonable alternative to conducting an RCT in this case, would be to perform what is essentially a retrospective cohort study on studies from a specific subdomain of medicine that differ according to the meta-level property in question. A standard retrospective cohort study is an

observational study where the difference in the risk of an outcome between two populations that differ according to some exposure is analyzed. Such a study in the context of research on meta-bias would be akin to one conducted by Hrobjartsson et al. (2012) on the effect of unblinded assessment in clinical trials.²⁸ This study compared the results calculated by blinded assessors with results calculated by nonblinded assessors by reviewing 21 trials which had both kinds of assessors. Hrobjartsson et al. (*ibid.*) found that effect sizes nonblinded assessors exaggerated effect sizes by 36% in comparison to blinded assessors. A similar approach can be adopted in research on the effects of meta-biases. Say we are interested in estimating the influence of publication bias on measurements of heart attack risk in statin research. This entails comparing the pooled estimate of the effect of statin therapy on heart attack risk in published studies to the pooled estimate of the effect of statin therapy on heart attack risk in all completed studies, published or unpublished. Here, instead of analyzing the outcomes of being exposed to a particular phenomenon, we are analyzing the outcomes of trials that share a common high-level property—in this case those trials that have been published versus those trials that have been completed.

While the retrospective cohort study model gives us an estimate of the influence of publication bias in practice, it suffers from a significant problem. The measures of effectiveness generated in clinical trials are not just subject to the problem of meta-bias, but other various challenges, from methodological biases to the encroachment of researcher subjectivity. Without a means to control for these problems, measurements of the effect of the meta-bias in question run the risk of being corrupted. A related problem is that meta-biases rarely occur in isolation—areas of research that are affected by publication bias are also affected by sponsorship bias—and thus, measures of the effect of one meta-bias risks being affected by another meta-bias. Controlling for these issues would require access to knowledge of the different phenomena at play in the various trials included in the meta-research study.

One plausible solution to this problem, proposed by Tabatabaei Ghomi and Stegenga (n.d.), is to simulate trial-level data using the higher order reported results of published trials, such as means and effect sizes. Importantly, using simulations allows researchers to not only specify the true effectiveness of the intervention in question prior to analyzing the effect of the meta-bias in question, but also control for methodological biases, researcher idiosyncrasies, and other overlapping meta-biases. Tabatabaei Ghomi and Stegenga (*ibid.*) have applied this methodology to research on antidepressants to measure the effects of publication bias on estimations of effectiveness.²⁹ In the study, different rates of publication bias are tested against different specified magnitudes of efficacy. The simulations show that publication bias has little effect on treatments

²⁸ Peter Gøtzsche (2013) argues that we should correct for lack of blinding in clinical research, using the results of this trial to demonstrate his proposed method.

²⁹ In their study, Tabatabaei Ghomi and Stegenga (2020) also simulate the effects of methodological aspects, particularly choices regarding categorization as a ‘responder’ to the treatment and the use of placebo run-in periods, on estimations of effectiveness in antidepressants research.

that were specified as highly effective but leads to overestimations for treatments that were specified as having low efficacy. While these simulations weigh different *possible rates* of publication bias, similar studies can be conducted using meta-research findings on the prevalence of various meta-biases. In doing so, we can measure their effects on different outcomes in different subdomains of medicine, and from there determine bias coefficients for these specific areas of research.

4.5 Conclusion

I began this Chapter by arguing that estimations of medical effectiveness are prone to failure because they do not account for the effects of what I refer to as meta-biases—aspects of medical research that lead to exaggerated research results, and which can only be detected by comparing sets of studies with particular meta-level properties. Meta-biases, I argue, distinct from methodological biases in that they are not directly linked to the methodological features of clinical research. Rather, they are connected to meta-level properties of a scientific discipline, such as the deep-rooted values of the medical research community or entrenched norms of the system in which medical research is conducted. I describe how meta-biases affect medical research and estimations of medical effectiveness by explaining two prominent exemplars of this type of bias, publication bias and sponsorship bias. Meta-biases lead to the systemic exaggeration of research results, which are then used to infer the general capacity of medical treatments (Stegenga 2018). As such, meta-biased results may lead to inaccurate estimations of medical effectiveness and indirectly affect extrapolation, making it difficult to assess whether the problem lies in measurement in clinical trials or evaluations of relevant differences between populations.

To remedy this, I offered a framework, which I call the bias dynamics model, for correcting measures of effectiveness in accordance with evidence on the effects of meta-biases. By employing this model, measures of effectiveness can be attenuated using empirical bias coefficients representing the effects of meta-biases. The bias dynamics model, and its use of bias coefficients, is appealing for at least two reasons. First, it has the potential to fill a gap in current evidence-based medicine guidelines, which emphasize the risks of methodological biases. In line with this, I argue that organizations like the Cochrane Group should publish regularly updated Second, bias coefficients provide a straightforward way for researchers to use up-to-date evidence to generate more accurate estimations of medical effectiveness.

Finally, I briefly outlined a provisional research program for determining bias coefficients. Extending on the work of Tabatabaei Ghomi and Stegenga (n.d.), I argue that simulations of patient-level data based on empirical results in published work can be used to measure the effect of meta-biases in clinical research. This proposal is in keeping with current trends in the relatively nascent field of meta-research. Such simulations should use evidence about the prevalence of meta-biases in specific subdomains of medicine to determine their effects on specific outcomes. The simulated results can then be used to generate bias coefficients for that area of medicine, which can be published on a regular basis to reflect the latest evidence on meta-biases.

Chapter 5

Medical Artificial Intelligence: What is Interpretability?

5.1 Introduction

Two sets of conceptual problems have gained prominence in theoretical engagements with artificial neural networks (ANNs). The first is whether ANNs are *explainable*, and, if they are, what it means to explain their outputs. The second is what it means for an ANN to be *interpretable*. In this Chapter, I argue that ANNs are, in one sense, already explainable and propose a novel theory of interpretability.

These issues often arise in discussions of medical AI systems (MAIS), where reliance on artificial decision making in medical contexts could have serious consequences. There is evidence that some of these systems have superior diagnostic and predictive capabilities when compared to human experts (Esteva et al. 2017; Fleming 2018; Rajpurkar et al. 2017; Tschandl et al. 2019). Indeed, many MAIS are already deployed in the clinic, including algorithms aimed at diagnosing retinal disease (De Fauw et al. 2018), and breast cancer treatment recommendations (Somashekhar et al. 2018) and screening (McKinney et al. 2020). For some, accomplishments like these are a precursor to the promising incorporation of machine learning (ML) into effective medical decision making (Wiens and Shenoy 2018). However, others have noted problems facing MAIS, including vulnerabilities to nefariously motivated adversarial attacks by various stakeholders in the healthcare system (Finlayson et al. 2019) and algorithmic racial biases in the prediction of future health care needs due to objective functions and training procedures (Obermeyer et al. 2019).

These problems have led to calls for MAIS, and ANNs in general, to be explainable—that is, if an ANN makes a recommendation, there should be an explanation for its decision (Athey 2017; Aler Tubella et al. 2019). In the context of healthcare, motivated by the need to justify artificial decisions to patients, some argue that maximizing the benefits of MAIS requires that their outputs be explainable (Watson et al. 2019). Others argue that having explainable MAIS makes it easier to recognize and remedy algorithmic rules which lead to inaccurate outputs (Caruana et al. 2015). On this view, having an explanation allows us to evaluate a MAIS' recommendations, which some argue is necessary for assessing the trustworthiness of such systems (see Ribeiro et al. 2016; Mittelstadt et al. 2019).

Issues of explainability have become problematic due to the prominent view that the accuracy of an AI system trades off against its explainability. Call this the *Accuracy-Explainability (AE) trade-off* (Gunning 2017; London 2019). Many of the problems MAIS are designed to solve are

complicated enough that achieving a high degree of accuracy requires highly complex ANNs. This, together with the assumption that highly complex ANNs are “less explainable,” leads to the peculiar reasoning that if an AI system being more accurate entails low explainability, and if low explainability entails an inability to assess trustworthiness, then a highly accurate AI system entails an inability to assess trustworthiness. However, this strikes as unintuitive since, presumably, accuracy should count towards trustworthiness.

There are three common strategies for dealing with this problem. Some argue that simplicity should be favored over accuracy when developing MAIS (Adkins 2017; Athey 2017). Others claim that we simply need to develop ways of increasing explainability (Gunning 2017). And for some, there is no problem at all since either we face similar issues when dealing with human decision makers (London 2019; Zerilli et al. 2019) or because simpler models may sometimes achieve the same degree of accuracy as more complex algorithms for the same task (Rudin 2019). Yet, all these approaches turn on what we mean by *explanation* and what features make explanations “good” or “fitting” for a given account. The literature on ANNs in general, and MAIS in particular, often makes use of the concepts of *explainability*, *understandability* and *interpretability*, but offer little critical engagement and contain persistent disagreement on the definitions of these terms (Lipton 2018; Krishnan 2019). Moreover, a central problem plaguing the ML literature is the conflation of these concepts: Explainability and understandability are typically treated as if they are synonymous with interpretability.

Several philosophers have analyzed these concepts more carefully. Zednik (2019), for instance, offers a pragmatic account of *opacity* to provide a normative framework detailing different kinds of knowledge that ought to be required by different stakeholders. Explanations, on this view, are the basis for how such knowledge is acquired. Creel’s (2020) account of *transparency* aims at illuminating recent successes in reducing algorithmic opacity and providing AI-explanations. Krishnan (2019) argues that pursuing definitions of interpretability, and related terms like explainability and understandability, is misguided because doing so reduces solutions to opacity-related problems to merely finding ways to make ANNs more transparent. Her claim that there are other solutions to these problems, while compelling, does not entail that defining interpretability in the context of ANNs should be wholly abandoned, particularly given the apparent conflation of terms. Páez (2019) recommends that focus should be shifted from explanation to understanding, arguing that traditional explanations of opaque systems are impossible and that understanding can be acquired by other means which do not require such explanations. Such strategies, Páez argues, exist in ML in the form of interpretive models aimed at understanding how an ANN functions, and post hoc interpretability aimed at understanding individual decisions. Taken together these accounts go some way to establishing interpretability as an important concept in its own right and its use within ML. However, none provide an explicit account of the term or how it should be used in analyses of ANNs.

The first aim in this Chapter is to clearly set the concepts of explainability, understandability, and interpretability apart. The explosion of different notions of ‘explanation’ in the context of AI has reinvented the wheel; philosophers of science have been developing notions of scientific explanation for nearly a century. I begin, in Section 5.2, by outlining four prominent accounts of explanation—the *Deductive Nomological*, *Inductive Statistical*, *Causal Mechanical*, and *New Mechanist* models—and argue that increasing the complexity of an explanation for some phenomenon does not make the phenomenon any less explainable. Then, in contrast to Páez’s (2019) claim that traditional explanations of ANNs are impossible, I argue in Section 5.3.1 that the four accounts I outlined in Section 5.2 indeed apply to neural networks, as they would to any scientific phenomenon. In this sense there is no AE trade-off. The source of much confusion within the literature is the conflation of the notions of explainability, understandability, and interpretability in cases where they are not interchangeable. Many claims within and surrounding the ML literature are explicitly lodged as calls for “explainability”, when it is the understandability of existing explanations that should be at issue. In Section 5.3.2, I briefly unpack the relationship between understanding and explanation, demonstrating that it is understanding that is defeasible by increasing complexity.

I then provide an explicit account of interpretability and offer a typology of interpretation methods in Section 5.4. I argue that interpretation is a relation between two explanations. During the process of interpretation, one explanation gives rise to a *more understandable* explanation. As with explanation, there are varieties of interpretation: *Total* or *Partial*, *Global* or *Local*, and *Approximative* or *Isomorphic*. This account of interpretability is consistent with many uses within AI, in keeping with philosophy of explanation and understanding, and provided with special attention to the accuracy-complexity relationship in MAIS.

5.2 The Indefeasibility of Explanation

In aiming for explainable AI (XAI) we ought to work with well-theorized conceptions of *explanation*. There are many accounts of scientific explanation; here, I employ those most germane to the problem of separating out “interpretability” as a special epistemic activity. While the recent uses of concepts of ‘interpretability’ and ‘interpretation’ are variously and sometimes inconsistently defined (see Section 5.4), “explanation” has a far longer and more rigorous conceptual history. Philosophers have reflected on the nature of explanation since before Socrates, but the modern discussion began in the late 1940s with Hempel and Oppenheim (1948).

I focus on the four models of explanation that have received the most significant attention: The Deductive Nomological model; the Inductive Statistical model; the Casual Mechanical model; and, more recently, the New Mechanist model. In this section, I briefly outline each of these models and then argue that explanations of each variety are indefeasible to complexity; that is, increasing the complexity of a phenomenon does not make it any less explainable. Establishing the indefeasibility of explanation is the first step to recognizing that the alleged trade-off between accuracy

(complexity) of ANNs and their explainability is misguided. Conceiving of this trade-off in terms of explainability has led to confusion in the debates surrounding interpretability and trust in MAIS.

5.2.1 Four Kinds of Explanation

Each model of explanation has the same broad structure. An explanation consists of an *explanandum*—the phenomenon being explained—an *explanans*—the elements of the explanation doing the explaining—and some *process of explanation* connecting the explanans to the explanandum (see Figure 5.1).

For *Deductive Nomological* (DN) explanation in particular, a successful explanation must satisfy two conditions (Hempel and Oppenheim 1948; Hempel 1965). First, the process of explanation must take the form of a sound *deductive* argument. Second, the explanans must have an essential *nomic* premise, i.e., at least one *law of nature* or *law-like proposition* without which the deduction would be invalid (see Figure 5.2). For example, the movement of a steel bearing on a flat table can be explained using the law, “all ferrous metals are physically attracted to magnetic fields” in conjunction with the fact that magnet has been brought close to the bearing.

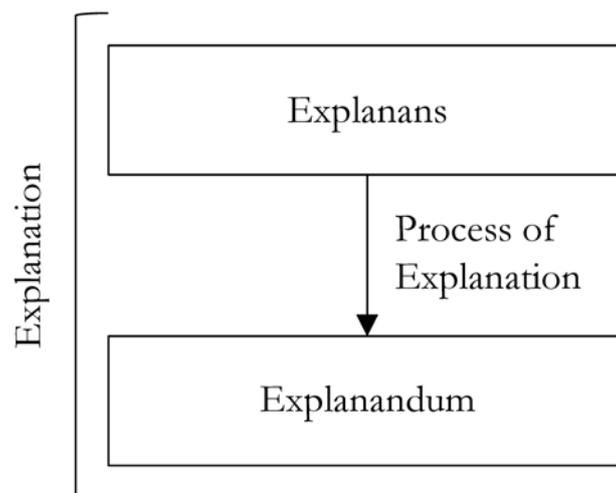


Figure 5.1. The general structure of explanation.

While the DN model is good for illustrating the explanation of phenomena which result from deterministic laws, it does not capture the characteristics of probabilistic events. In response to this, Hempel (1965) introduced *Inductive Statistical* (IS) explanation. IS explanation involves the inference of an individual event from a *statistical law* and empirical information about the event (see Figure 5.2). For example, the increased probability of having breast cancer given a mutated *BRCA1* gene in conjunction with a particular patient having a mutated *BRCA1* gene explains the patient having breast cancer. Here, the relation between the explanandum and the explanans is inductive because all that can be inferred from the information given is there being a higher or lower probability that the patient has breast cancer. If the probability of having breast cancer given that the patient has a mutated *BRCA1* gene were lower, then even if the patient has breast cancer, this

information cannot be used to explain it. The basic idea then is that the success of an IS explanation depends on whether the explanans entails a higher probability of the explanandum obtaining.

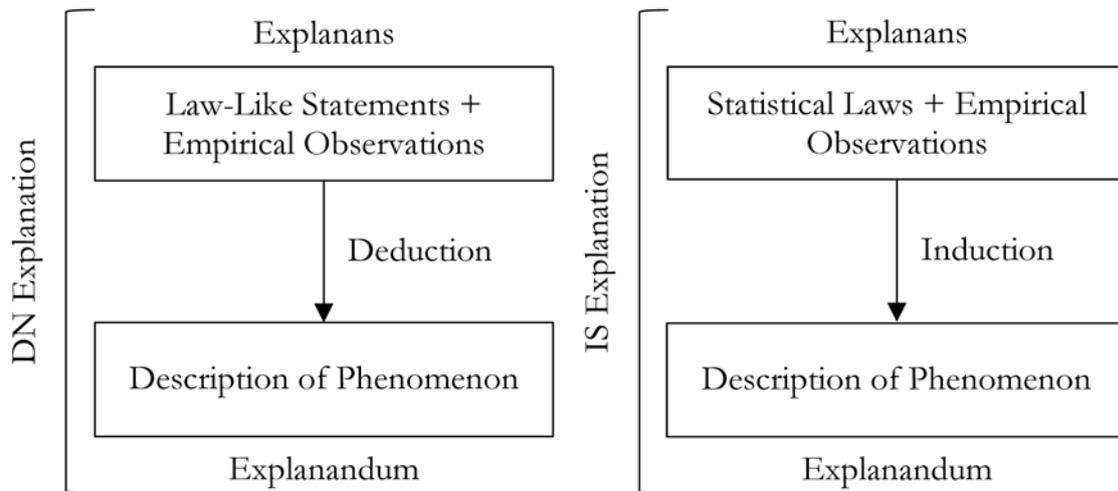


Figure 5.2. The structures of DN and IS explanation.

Two decades after Hempel’s work on the DN and IS models, Salmon (1984) proposed the *Causal Mechanical* (CM) model, which he claims highlights the central role of causation in explanation.³⁰ The idea behind the CM model is that explanation involves showing how the explanandum fits into the causal structure of the world. There are two important aspects of causation which feature in CM explanations. The most basic of these, a *causal process*, is the ability to transfer a *mark* or its own physical structure in a spatiotemporally continuous way. An example of this would be the movement of sound waves through air, or the movement of a particle through space. The second, a *causal interaction*, occurs when causal processes interact with one another resulting in a modified structure. Examples include sound wave interference, or the collision of particles. A successful CM explanation involves citing some parts of the causal processes and causal interactions which lead to the phenomenon in question (see Figure 5.3). We might, for instance, explain the phenomenon of noise cancellation by citing the existence of two sound waves, one with inverted phase to the other (the causal processes) interfering with one another (the causal interaction).

A cluster of views of explanation has recently emerged, all termed *New Mechanist* (NM). These views all center on the idea that providing a mechanism is important for explanation, and originate largely from the work of Machamer et al. (2000), Bechtel (2011), and Craver and Darden (2013). The idea behind NM accounts is that providing an explanation involves showing how some

³⁰ Prior to this, Salmon (1971) proposed his statistical relevance model of explanation. Although we do not cover it here, it is worth noting that he does not wholly abandon the notion of statistical relevance, but rather claims that one typically acquires information about statistically relevant relations before providing a CM explanation.

phenomena arise from a collection of *entities* and *activities* (see Figure 5.3).³¹ A successful explanation involves identifying the entities and activities that bring about a phenomenon with regularity and without gaps, missing entities or activities. For example, an explanation of protein synthesis will need to identify the entities involved (ribosomes, aminoacyl-tRNAs, mRNAs, GTP, etc.) and the activities (binding of mRNA to the ribosome, hydrolysis of GTP, translocation of mRNA within the ribosome, etc.), which are typically *depicted* in the form of a mechanistic diagram of the sort familiar from textbook mechanisms. A mechanism also includes two special collections of entities and activities, the initial or “set-up” and the final or “termination” conditions, in this case the binding of mRNA to the ribosome and the final release of a synthesized protein respectively. However, to better describe biological mechanisms exhibiting cyclic organizations, some argue that such conditions are not necessary for mechanistic explanations (Bechtel 2011). Yet, since ANNs are never cyclic, Machamer et al.’s characterization of mechanisms is apt for explaining such systems. Moreover, when ANNs are recurrent, the pathways can always be unrolled and mechanistically depicted.

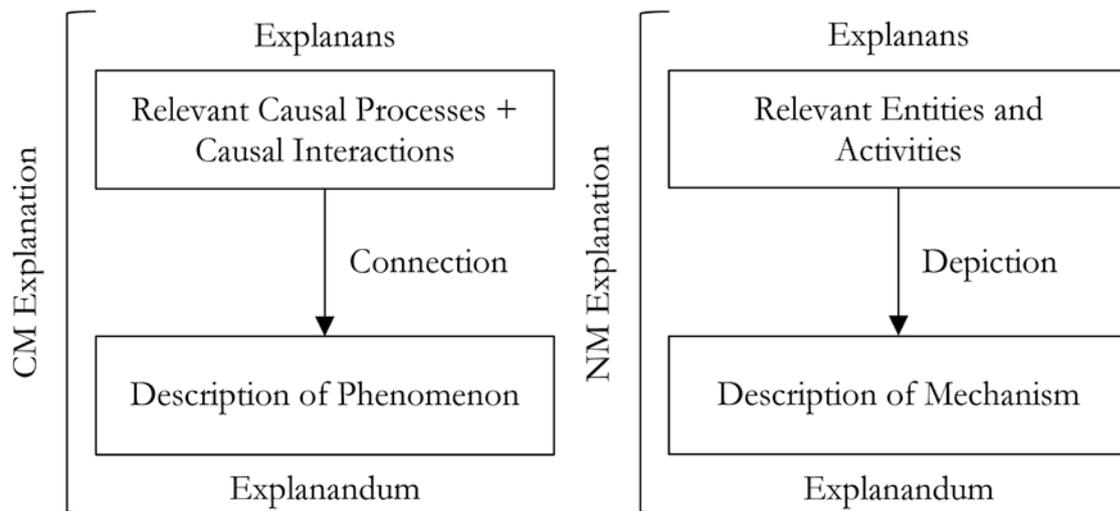


Figure 5.3. The structures of CM and NM explanation.

There are several criticisms of the accounts of explanation described above (see Salmon 1989; Skillings 2015; Godfrey-Smith 2016). While I do not have the space to engage with all of these, there are two that may have implications for the arguments that follow, and are thus worth addressing. First, it may be argued that some of the typically cited challenges facing DN, IS, and CM explanations, such as explanatory asymmetry and explanatory irrelevance, are the result of disregard for, or inadequate treatment of the role of causation in explanation (Woodward 2003). Woodward’s (2003) causal-interventionist account of explanation, according to which successful explanations

³¹ As Levy (2013) notes, advocates of the explanatory value of mechanisms are “fairly coy” (p. 102) about defining the key notions, a matter further complicated by a lack of universal terminology. E.g., sometimes “parts” are discussed instead of “entities” (Bechtel 2006). However, for present purposes only the outlines of the NM approach are required.

exhibit patterns of counterfactual dependence, may help deal with these issues, particularly when it comes to the problem of explanatory relevance. Since I argue in Section 5.3 that ANNs are explainable using the traditional models described above, it may be argued that our account of interpretability could inherit some of these issues. I agree that the causal-interventionist account may be useful for solving some questions about explainability, understandability, and interpretability in AI (Páez 2019). Indeed, I describe interpretability processes in Section 5.4.3 some of which result in understandable explanations which include counterfactual information, however, the causal-interventionist account includes pragmatic elements that we maintain should not count against a phenomenon's *explainability*, namely that an explanation should be "deeper or more satisfying" (Woodward 2003: 190) than those provided by DN, IS, or CM explanations. Of course, I agree that some explanations can be more preferable than others given some explanatory virtue, such as simplicity, but we disagree that such virtues should be considered when evaluating the *explainability* of a phenomenon.

Second, proponents of the ontic conception of explanation, according to which explanations are physical composites of the entities being represented, may argue that the characterization of NM explanation above is misguided. Explanations, on this view, are not *depictions* or *representations* of the mechanisms which constitute the phenomenon, but rather physical entities themselves which actively contribute to the causal structure of the world. Accordingly, there is "no question of explanations being 'right' or 'wrong,' or 'good' or 'bad.' They just are." (Craver 2007: 27). The ontic conception supports my position regarding the indefeasibility of explanation below. It does so, however, in a problematic way. On a strongly ontic conception of explanation, there *is* an objective explanation of ANNs, provided only that there *are* ANNs. Since there are ANNs, they would be automatically objectively explainable on this conception of NM. But this purchases the indefeasibility thesis at the cost of relevance to issues of explainability in ML. Calls for XAI are not, to my mind, calls for objective explanations.

I agree with Craver (2007) that what makes something an explanation is not some set of pragmatic conditions such as whether it produces understanding, or whether it is a "good" explanation because it satisfies some set of explanatory virtues, but rather whether it accurately maps onto the world via any one of the explanation processes outlined above. My concern with explanations and use of the concept of explanation are with "explanatory texts" (Craver 2007), that is, how explanations are given by scientists, ML researchers in particular. Nonetheless, the rejection of the involvement of pragmatic conditions does not entail or require the strong ontic thesis above. Even if one finds an explanation undesirable because it does not satisfy some particular set of explanatory virtues or pragmatic conditions, that does not make it any less an explanation. And if there is an explanation for a given phenomenon, then that phenomenon is explainable.

5.2.2 *The Indefeasibility Thesis*

Unfortunately, the well-developed typology of explanation described above is largely neglected in discussions of explainability in neural networks. Those who argue for the inverse relationship

between complexity and explainability for ANNs have overlooked an important feature of explanation as it is understood above: its *indefeasibility*. They often treat complex MAIS as though they had outgrown their explainability, but this is not quite right. In fact, the explainability of a phenomenon— such as the output of an AI system—depends only on the relationships between our background theories and evidence, and the procedures of the explanation process used. In other words, in each of the accounts of explanation introduced above, the features that make something an explanation turn out to be invariant with respect to the complexity of both the explanans and the explanandum. What the considerations in this subsection illustrate is the following *indefeasibility thesis* for explanation of the phenomena characteristic of ANNs: Adding complexity to the explanation of the phenomenon does not entail that the phenomenon is any less *explainable*.

This is not a claim about the quality, superiority, or goodness of a given explanation. The concern is whether increasing the complexity of a given explanation makes it no longer an explanation. While increasing complexity may reduce the quality of the explanation, by making it less preferable or understandable (see Section 5.3.2), or even in some cases make it a bad explanation, it does *not* make it any less *an explanation*, and thus the explainability of the phenomenon of interest is unaffected. Let me clarify this position by looking at each kind of explanation in turn.

Consider a DN explanation of some phenomenon beginning with a simple explanans, a set of law-like premises \mathbf{L} and empirical observations \mathbf{E} , that concludes with the explanandum/description of a phenomenon x after a valid deduction. What makes this a DN explanation is just that the explanandum is reached deductively using an explanans that is partially law-like. Now consider what happens if we expand the explanans to include more premises, which are consistent with our best background theories and evidence, say by substituting one of the law-like premises for a set of n laws jointly entailing it. This may make the explanans less manageable or more difficult to understand, but we still have a DN explanation. That is, if x follows deductively from $\mathbf{L} + \mathbf{E}$, then it also follows from $\mathbf{L}^* = \{\mathbf{L}_1 + \mathbf{L}_2 + \dots + \mathbf{L}_n\} + \mathbf{E}$, where \mathbf{L}_{1-n} are (complexity-increasing) laws which, taken together, entail \mathbf{L} . The expansion of the explanans does not make its laws any less law-like nor its conclusion any less deductively valid. Put another way, the explanandum is no less explainable, since the connection to the explanans, no matter complexity, is still deduction.

We might also increase the complexity of a DN explanation by substituting the empirical observations with a more complex set of observations, say by describing them at a more concrete level of abstraction, or with more accuracy. To account for how these observations play a role in explaining the phenomenon of interest, we may require a more complex set of laws. Suppose we have the explanans $\mathbf{L} + \mathbf{E}$ as before, and consider some more complex set of true laws \mathbf{L}^* and/or more complex set of empirical observations \mathbf{E}^* from which x follows. Since \mathbf{L}^* is a set of laws (or at least law-like propositions), and \mathbf{E}^* is a set of empirical observations, the deduction of x from this (more complex) pair constitutes a DN explanation of x . As above, the explainability of x does not depend on the complexity of its explanans, but on whether the relationship between explanans and explanandum is deductive. To illustrate, consider the example of the steel bearing moving on the flat

table from before. We might increase the complexity of the explanation by noting more empirical observations, such as the initial distance of the magnet from the bearing, surface friction forces, and so on. Or we may explain the movement of the steel bearing by using more detailed laws including those governing the total induced magnitude of force on the ferromagnetic material's volume, and laws of motion, among others. These laws, taken together, entail that ferrous metals are attracted to magnetic fields. While this more complex explanation may be more difficult to understand and is by most accounts less appealing than the simpler one, it is still a DN explanation for the movement of the steel bearing on the flat table.

Importantly, not all ways of *changing* an explanans are only additions of complexity, some changes affect the quality of the explanation produced. For example, we might change a satisfying DN explanation of the dissolution of salt to append adjunct claims about the dissolution of “hexed salt” (see Kyburg; Salmon 1989). In doing so, we included irrelevant information, since the hexing presumably plays no part in the dissolving of salt. However, we do not defeat a DN explanation by conjoining it with adjunct, irrelevant premises—though perhaps we worsen it markedly. Note, if we had changed the explanans to say only that “all and only hexed salt dissolves” and removed claims that “all salt dissolves,” then we *would* have failed to provide a DN explanation—since that assumption is inconsistent with our best background theories and evidence. That sort of change is not an addition of complexity, it is a substitution of logically distinct premises. Moreover, the complexity introduced in highly accurate ANNs is not irrelevant in this way; there is no worry that explainability is defeated by magical or irrelevant suppositions about complex MAIS.

The situation is somewhat different for IS explanations. Suppose we have an IS explanation of some phenomenon starting with statistical laws **S** and empirical facts **E** and ending with some inductively inferred conclusion as its explanandum, which might be some particular outcome x or a probability of such an outcome $P(x) = p$ (see Woodward 1989, 2019). Now assume we discover that some other statistical laws or empirical facts **C** are relevant to the phenomenon in question. Notice that additional statistical laws or empirical facts will, under certain conditions, affect (the probability of) the explanandum. However, our aim is not to show that increasing the complexity of the explanans does not affect the support for the explanandum. That is, if x follows inductively from **S** + **E** with support p (or $P(x) = p$ follows inductively from **S** + **E**), then x follows inductively from **S** + **E** + **C** with support $p \pm c$ (or $P(x) = p \pm c$ follows inductively from **S** + **E** + **C**).³² What is important here is that despite increased complexity, the process of explanation is the same, so the relationship between the explanans and explanandum remains one we are happy to call explanation. Of course, the explanandum may thereafter be explained with low probability, making the explanation less “successful” or “good,” but not making the explanandum less explainable by the inductive process of IS explanation.

³² There are cases where the addition of an empirical fact may lead to a deductive explanation, for our purpose we can treat this as a trivial case of inductive explanation, where the probability is zero or one.

Similarly, increasing the complexity of the explanans in CM explanations does not make the explanandum any less explainable. Consider a CM explanation which appeals to a particular set of causal processes \mathbf{P} and set of causal interactions \mathbf{I} which together bring about phenomenon x . We might discover additional causal processes and/or causal interactions, which are relevant to x (say by uncovering finer grained details about the particular causal process) such that a more complex set of causal processes \mathbf{P}^* and/or set of causal interactions \mathbf{I}^* together bring about x . While these new sets of causal features may make the explanation harder to follow, they do not make the CM explanation any less an explanation than before. In other words, just because $\mathbf{P}^* + \mathbf{I}^*$ is more complex than $\mathbf{P} + \mathbf{I}$, does not make x any less explainable by the former than it is by the latter. Similarly, for an NM explanation. Such an explanation will depict relevant entities \mathbf{E} and activities \mathbf{A} as responsible for the phenomenon x in question. While the discovery of additional entities \mathbf{E}^* and/or activities \mathbf{A}^* may make the explanation of x less intelligible, it does not make the phenomenon any less explainable than before. Complex mechanisms are mechanisms no less.

Indeed, there may be system dynamics of some phenomena, such as a high degree of stochasticity (Skillings 2015; Godfrey-Smith 2016), that make the discovery and articulation of mechanisms or causal processes difficult, or proposed mechanisms dissatisfying. Nonetheless, granted some existing mechanism or process, adding complexity to it in the form of entities and activities, even stochastic activities, does not destroy the mechanism. Similarly, some may worry that complex ANNs are nondecomposable, which is to say, they are impossible to (mechanistically) explain because they are so hierarchically complex that we cannot assume that the interactions between the relevant entities and their activities can be described linearly (see Bechtel and Richardson 2010). However, even if concerns surrounding nondecomposability are warranted in other domains, they do not apply to ANNs since such systems, as we argue below, can be described in NM terms.

Together, these considerations establish the indefeasibility thesis as stated above. One may justly worry that this thesis has been bought at too hefty a price, since they are asked to admit many “bad” or “dissatisfying” cases as explanations. In part, this is just the price of doing science, where many explanations, perhaps even many we find good or satisfying today, turn out not to live up to the standards of good explanation. Of particular interest in this connection is when an explanation is deemed bad or dissatisfying because it does not produce understanding. Happily, some of the procedures of science in general and ML in particular aim directly at remedying this by producing better or more satisfying explanations from those less so. I describe and explicate these procedures in the context of ML below (Sections 5.4.1–5.4.4), collectively under the heading of interpretability methods. Some of these bad explanations are remedied in the course of good science. For now, I turn to the specifics of applying these varieties of explanation to MAIS, confident that the complexity added to MAIS is immaterial to explainability, even if dissatisfying.

5.3 The Explanation in the Machine: Medical AI Systems

Having established the indefeasibility of explanation, we can apply it to the explainability of MAIS. In this section, I argue that even extremely complex MAIS can be explained using each of the models of explanation outlined above. That is to say, we are always able to generate an explanation of an AI model based on particular features of the model itself. The upshot here is that the AE trade-off problem is in fact not one of explainability—we do have access to explanations of MAIS. Each account of explanation applies to ANNs as it would to any other scientific phenomenon of interest, what differs is our capacity to *understand* these explanations. The precise problem is the difficulty of understanding the explanations of complex ANNs. This is because, in contrast to explanation, understanding is defeasible. The AE trade-off problem should therefore strictly be cast as an issue of providing *understanding* of complex AI systems. Conveniently, a plausible solution to this problem has already been proposed: interpretability. A further problem, however, is that there is little agreement over exactly what interpretability is, an issue I deal with in Section 5.4.

5.3.1 *Medical AI Systems Are Explainable*

It is possible to explain MAIS using each of the four models outlined above. What follows are explanation-sketches (see Hempel 1965) of how ANNs can be explained using each model. To illustrate, take a recent deep learning MAIS, developed by MIT and MGH researchers, which is designed to identify and assess dense breast tissue, an independent risk factor for breast cancer. An associated study shows that this system closely matches the reliability of expert radiologists in evaluating breast tissue density and that the MAIS has been successfully deployed in a clinic setting (Lehman et al. 2019). It is built on a ResNet-18 convolutional neural network (CNN) used for computer vision tasks.³³ The model was trained and tested using a dataset of 58,894 randomly selected mammograms from 39,272 women between January 2009 and May 2011. Of these, 41,479 were used for training and 8,977 for testing. In the testing phase, the MAIS matched with expert assessment at 77% across a standard Breast Imaging Reporting and Data System (BI-RADS), which categorizes tissue in four ways: fatty, scattered, heterogenous, and dense. In binary tests, either dense or non-dense tissue, the model matched expert assessments at 87%. Subsequently, the MAIS was deployed for testing in a clinical setting at MGH and performed well, matching with expert assessments across all four BI-RADS categories at 90% and at 94% for binary tests. The basic idea is that the MAIS takes inputted images and outputs a classification according to categories for breast cancer risk.

To understand how this MAIS can be explained, we must know a little about its system architecture. The ResNet-18 consists of 18 layers. The first 17 layers are all convolution layers, and

³³ A CNN is an ANN particularly adept at detecting patterns in images. It uses convolutional layers to split an image up into overlapping tiles, which are then analyzed for learned patterns before the signal is outputted to subsequent layers (see Goodfellow et al. 2016, Chapter 9).

the last is a fully connected network. This type of ANN model is most commonly used to classify images—it takes an image and outputs a classification.³⁴ This is done by feeding the image to the ANN’s input layer as a set of numbers which then get processed through the ANNs weighted nodes and edges and outputted as a vector of real numbers. Typically, the image will then be assigned to the most probable class.

The DN model can be used to explain any particular image classification produced by this MAIS. A DN explanation of how the MAIS assesses an input image involves listing the weights attached to each and every node and the informational routes indicated by each and every edge at every convolution stage, and the weights of the fully connected network along with the assigned numerical values being fed into the input layer and the network architecture. Once we have that, we can list the values for the classifications the MAIS learned in the training and testing phases of development, and see that its classification of the image is based on comparing the ranges of these classifications with the output value of the image. In doing so, we are explaining the explanandum—here, the MAIS classifying of image I as classification c —using an explanans consisting of a law-like premises—in this case, how the weights of all relevant nodes and edges produced the output value, along with the law that an output is assigned to the most probable class—and additional information about I —which includes the set of input values assigned to I , and the output value c .

While the DN model gives us an explanation of specific classifications, the IS model can help explain the probability that the MAIS’s classifications are accurate. Broadly, the accuracy of the outputs of ANNs can be explained by appealing to details about the training process as statistical laws and the nature of the training data used as empirical information.³⁵ If these, taken together, inductively entail a probability that the ANN’s outputs are accurate, then we have successfully explained the accuracy of the ANN’s outputs. In the case of the MAIS above, we can explain its high degree of accuracy, here the matching with expert assessments, by citing the training procedure and details about the mammogram image dataset used.

For CM explanations of ANNs, we would cite the causal processes and causal interactions involved. This would entail describing the ANN using oft-used terms of art, drawing on biological analogies. ANNs are constituted by connected *nodes*, sometimes referred to as artificial neurons that, like biological neurons, receive and send information, or *signals*, via connections resembling biological synapses, termed *edges*. Signals are typically approximations of real numbers that, when received by nodes as *inputs*, are processed using some non-linear function of their weighted sum, and sent as *outputs* to other nodes. Nodes typically have weights (and biases), which are tuned when the ANN is trained. These weights control the strength of output signals and can, in the case of

³⁴ While we use an example of a diagnostic MAIS, viz. a classifier, our arguments apply to ANNs designed for other goals, such as automated planning and personalized recommendations.

³⁵ Karimi et al. (2020) deploy an approach that allows assessment of the classifications of ANNs on the basis of causal information associated with the generation of training data. Provided such causal information is obtained by statistical assessment of the target for MAIS classification, it can figure in either IS or CM explanations of ANN accuracy.

nodes with thresholds, determine whether a signal is sent at all. The nodes in ANNs are typically organized into *layers*, which can perform different tasks by processing their inputs in different ways. Two layers are special: the *input layer* where information is fed into the ANN and the *output layer* which returns a result. Typically, each of the output nodes provides a probability that the input belongs to a particular class. The connections between nodes are essentially functions of the data that is fed into them. Information is fed into the input layer and passes through the different layers until it reaches the output layer, producing a classification. This sort of description can be applied to particular ANNs like the MAIS above. Feeding the mammogram image into the input layer of the ResNet-18 and the convolutional operations of each layer are causal processes, while the signals sent between the different nodes and layers are causal interactions. These causal processes and interactions lead to the output of the MAIS.

For NM explanations of ANNs one has a choice between levels. We can explain the phenomenon of classification at a high level of abstraction, including in our mechanism only the coarse input-output depiction of an ANN; at the level of the abstraction of nodes, signals, and edges; or delve into the details of a mechanistic explanation at the level of the hardware circuitry involved in computation. Jacobson (1959), for example, shows how circuit-board diagrams generally can be abstracted into simpler networks of nodes and connecting edges, and the same process can, in principle, apply to any ANN. A circuit or hardware-level mechanistic depiction requires specifying the relevant entities and activities involved and depicting how their arrangement and interconnection results in the computation of a classification output from inputted image data. Arguably though, this sort of reduction to electrical signals and processes of computation is not the ideal level of mechanistic explanation. Machamer et al. describe the importance of selecting the right level of explanation in terms of “bottoming out” of mechanisms—describing mechanisms typically ends at a level “accepted as relatively fundamental or taken as unproblematic” (2000: 13) by a given discipline. In the context of explaining MAIS classification, we can presumably treat hardware processes as elementary and focus on ANN architecture. At least, until there is some indication that features of MAIS classification turn on lower-level mechanisms, the explanation ends there.

Just as with CM explanation above, the essential features of an NM explanation relevant to MAIS will involve identifying entities (nodes and layers) and activities (signaling, input and output), start-up conditions (input data and configuration of the network) and termination conditions (output of and thresholds for classification), the essentials of which are described above. Indeed, there may also be “higher level” mechanistic explanations, which identify more intuitive activities of specific layers or nodes, and these might be more easily “understood” or “visualized” than the mechanistic depiction of the architecture of ANNs (see Section 5.4.1, and Lipton (2018) on decomposability). But the existence of higher-level explanations does not prevent NM explanation of MAIS at the level of ANNs themselves.

It is important to distinguish between whether a mechanism is satisfactory or good at some level of abstraction, and whether it is a genuine NM explanation. There is justifiable concern about

whether a given NM explanation of an ANN is a good one, particularly when that ANN itself is treated as a model of some other phenomena we are interested in explaining (e.g., the mammalian neocortex, see Buckner 2019). But the provision of an NM explanation must of course precede assessment of its quality. Moreover, MAIS are in some ways actually simpler than the usual targets for NM explanations—biological neurons—since the NM explanations of MAIS in their current form need not account for chemical or analogue features of artificial neurons.

The above discussion illustrates that explanations of ANNs are available to us. The AE trade-off, posed as an issue of “explainability” as it commonly is in ML literature, is therefore not as problematic as one might think.

It is worth considering a possible challenge to this position. Some have argued that the traditional accounts of explanation outlined above are incomplete because they do not take pragmatic elements into account (e.g., van Fraassen 1980; Achinstein 1983). For instance, it has been suggested that explanations are only *good* or *successful* when they provide understanding, and that explanations should exhibit a proper relationship to their audience (Potochnik 2016). Consequently, some may contend that my view of explanation fails to adequately account for the context or subject for which a given explanation is intended. According to this objection, the explanations of ANNs provided by the traditional models outlined above should not really be considered explanations at all since they do not guarantee understanding of the outputs given to us by AI systems. While my position is bound to be contentious to some ML researchers and philosophers favoring pragmatic accounts of explanation, there are several reasons for preferring it over such accounts—particularly in efforts to disentangle interpretability from explainability.

First, my view is compatible with widely accepted pluralism about explanation. Pragmatic accounts are also pluralist, but part of the pragmatist argument against traditional accounts is that these assume some overarching account of explanation which covers all contexts, but this is misguided. When explaining ANNs, we can use the account most suited to our given aims. That is, when providing explanations of these traditional sorts, we can account for features of explanatory contexts without adopting a pragmatic account of explanation itself.

Second, these traditional models of explanation share a common structure (Figure 5.1) that is helpful in defining interpretation. Indeed, since pragmatists do not dispute this common structure but add to it, the general features of the proposed account of interpretation are adaptable to pragmatic accounts of explanation. Only, they will need to employ more conceptual machinery than necessary to provide an analogous account of interpretability.

Third, and most important, one of the aims of this Chapter is to argue for the value of separating explanation and understanding in the context of XAI. That is not to say that these concepts are wholly separate. In the next subsection, I contend, in line with much recent work in philosophy of understanding, that explanation and understanding are indeed related, just not as strictly as many ML researchers and proponents of pragmatic accounts think. Further to this, setting these notions apart demonstrates that the problem of complexity really lies in its tendency to trade off against

understandability. This is crucial to developing an account of interpretability which successfully describes many of the methods used by many ML researchers for increasing understanding of ANNs.

5.3.2 *Separating Explanation and Understanding*

Before developing an account of interpretability, it is worth parsing part of the relationship between explanation and understanding. There has recently been a surge of philosophical interest in the concept of understanding (see de Regt et al. 2009; Strevens 2011, 2013; Potochnik 2016; de Regt 2017; Khalifa 2017). Although a full characterization of the concept is well beyond the scope of this Chapter, these existing accounts illuminate the important differences between understanding and explanation that illustrate the defeasibility of understanding.

Some contemporary accounts of science affect somewhat of a revolution by switching focus from explanation to understanding. Because of this, some may be tempted to align them with the pragmatic notions of explanation referred to previously. Common to these accounts, however, are the claims that (1) understanding necessarily involves *having an explanation* and (2) understanding demands satisfying some other condition(s) which are *not dependent on the qualities of the explanation alone*.³⁶ Though, there is little consensus about what these other conditions are. Most agree with Potochnik (2016) that this will involve some relationship between the explanation and the explainer or audience, but disagree about what relationship is required. For instance, de Regt argues that, “A phenomenon P is understood scientifically if and only if *there is an explanation* of P that is based on *an intelligible theory* T and conforms to the *basic epistemic values* of empirical adequacy and internal consistency” (2017: 93, my emphasis). Strevens (2013) argues that understanding involves “grasping” *a correct scientific explanation* and characterizes this in terms of possessing a particular *psychological state*. Khalifa (2017) agrees that one understands a phenomenon to the extent that they “grasp” an “explanatory nexus”, adding that grasping refers to a *cognitive state* resembling scientific knowledge.

In stating that an explanation is necessary for understanding, condition (1) above helps illustrate the untenability of the claim that if something fails to give rise to understanding, then it is not an explanation. Accepting, in line with much current work on understanding, that (1) is true amounts to accepting that if you understand some phenomenon then you can explain it, or contrapositively that if you cannot explain some phenomenon then you do not understand it. It would be false to conclude from (1) that if you do not understand some phenomenon then you cannot explain it. Simply put, you can explain things you cannot understand—doing just this is a part of the learning process and perhaps a psychological preliminary to understanding generally—you just cannot understand things you cannot explain.

³⁶ A notable exception is Lipton (2009), who argues that we can have understanding without explanation. However others, including Strevens (2013) and Khalifa (2017), have disputed this.

Where explanation and understanding are set apart is in terms of (2). Whatever particular view of understanding one may prefer—whether it is “grasping,” “intelligibility,” “apprehending,” or “knowing”—it is, at least in part, subjective or contextual. The intelligibility of a scientific theory, which is necessary for understanding in de Regt’s account, is by his own lights dependent on a scientist’s being able to “recognize the qualitatively characteristic consequences of T without performing exact calculations” (2017: 102; also see de Regt and Dieks 2005). What this means is that intelligibility, and thus understanding, partially relies on subjective features of the individual who is trying to understand the phenomenon in question. For both Khalifa and Strevens, grasping an explanation, and thus whether a given explanation actually provides understanding, will turn on psychological features specific to the user of that explanation, for example, on features of the scientist, engineer, doctor or patient. Plainly, what is understandable to Mandy might not be to Paul; what is understandable to the engineer of an ANN may not be to a radiologist or the person using a MAIS.

Taken together, (1) and (2) show two things: First, the pragmatic objection to the explainability of ANNs is misguided, thus strengthening the argument that explainability is guaranteed by the sorts of things that appear in the explanans and their relationship to the explanandum. That is why explainability is infeasible by increasing complexity: The properties of the explanans and relationship to the explanandum remain the same come whatever complexity.

Second, because one’s understanding is conditional on subjective features (whether psychological, cognitive, or contextual), it is not impervious to complexity. The capacity to understand varies between individuals and is sensitive to the complexity of both the explanans and the process of relating it to some explanandum, phenomenon. Understanding, in contrast to explanation, *is* defeasible. If ANNs are strictly speaking explainable but often those explanations are not understandable, then what is needed is methods for making those explanations more understandable. This is precisely the reason that we need to move beyond mere individual explanations to interpretation.

5.4 Interpretability

Much confusion in the debate about, and push for, XAI can be attributed to a conflation of explainability and interpretability. Interpretation has been devised and variously defined within the sciences themselves (Ribeiro et al. 2016; Mittelstadt et al. 2019; Lipton 2018; Krishnan 2019; Páez 2019). Philosophers and scientists are not at fault for misunderstanding the scientific notion of interpretation since there is no single such notion on which to rely. Indeed, interpretability is often connected directly with explanation and understanding, although by way of unhelpful equation or equivocation. Lipton (2018) says of interpretability that it “reflects several distinct concepts,” which is to say that it is used inconsistently, or at best equivocally. Indeed some, like Miller (2019, 8), are happy to accept equating interpretability with explainability. In contrast, I argue that it is best to set these concepts apart. I begin with an account of interpretation, from which an account of interpretability follows straightforwardly as the ability to provide an interpretation.

5.4.1 What is Interpretation?

Fundamentally, *interpretation is something that one does to an explanation to make it more understandable*. When we do not find an explanation understandable, we ask for it to be interpreted. We can also talk of “interpreting data,” “interpreting a phenomenon,” or “interpreting an ANN” but when these do not mean precisely the same thing as “explaining data,” “explaining a phenomenon,” or “explaining an ANN” then these uses are derivative of the notion of interpretability proposed here. To “interpret data” is to do something to an explanation of/involving data; to interpret an ANN is to do something to an explanation of/involving the ANN. An interpretation is also something that is performed on each essential part of an explanation: one interprets an explanans, process of explanation, and explanandum.³⁷ An interpretation then gives rise to *another explanation*, by one or a combination of the current techniques of XAI, or interpretability methods detailed below (Sections 5.4.2–5.4.4). Finally, when successful, an interpretation gives rise to an explanation that is, in some way or another, *more understandable* than the explanation we began with.

Take the example of the breast tissue classifier from before: We might start with a traditional explanation, such as a DN explanation, of a particular output from the MAIS. Given the complexity of such an explanation, it may be extremely difficult to understand for most users; thus, one may ask for a more understandable explanation. We can then apply one of the interpretability methods with the aim of producing a new, more understandable explanation. To generate such understanding, one might restrict an explanation to only that information pertaining to how an individual patient was classified (local interpretation Section 5.4.3 below). In doing so, we obtain a new explanation from the old, although one presumably less complex—since localized to a single case—and ideally thereby more understandable.

Terminology about interpretation methods often differs without changing the overall concept. Zednik, for example, identifies several techniques to “render opaque computing systems transparent” (2019: 8), such as input heatmapping and feature-detector visualization. Insofar as transparency involves production of explanations that are more understandable, these are simply cases of interpretation (specifically local and approximate interpretation Sections 5.4.3–5.4.4). Watson and Floridi (2020) offer a formal framework, which they refer to as an “explanation game,” by which proposed explanations are scored on three properties: *Accuracy*, *simplicity*, and *relevance*. Accuracy refers to how precisely the explanation models the ANN’s processes from input to output. Simplicity is equated with the understandability of the explanation for a given stakeholder. And relevance refers to whether the elements cited in the explanation are indeed responsible for the output being explained. On this view, the goal of interpretable machine learning is to provide explanations that maximize the balance between accuracy, simplicity, and relevance. Given that the

³⁷ Although in real cases, many of these components of interpretation may be implicit or trivial, i.e., one may leave the explanandum unaffected (see Section 5.4.2 below on partial interpretations).

explanation game aims at producing explanations that are more understandable for a given user, I take this framework to be describing cases of interpretation (specifically total and approximate interpretation Sections 5.4.2 and 5.4.4). Indeed, that there is something like a connection to the production of greater understanding is perhaps the only consistent feature across accounts of interpretability on offer.

Lipton (2018) and Krishnan (2019) make it clear that the notion of interpretability is often undefined or presumed intuitive and implicit by scientists and engineers. The choice to characterize interpretation at the level of explanations themselves is in part an attempt to remedy this by providing a workable definition of the *process* of interpretation. Furthermore, the view of interpretation proposed in this Chapter does capture some of the popular formulations of the notion offered within the ML community. This account of interpretation attempts to strike a balance between consistency with, and unification of, scientific usage.

Consider the view of Ribeiro et al. on the explanation of ANNs: “An essential criterion for explanations is that they must be **interpretable**, i.e., provide qualitative understanding between the input variables and the response” (2016: 1136). One way to read this is as asserting that interpretability is merely synonymous with understandability; another is to see it as a special case. If an explanation is already understandable, then surely it is interpretable, since it gives rise to an understandable explanation (itself), trivially. But, as we have noted, not all interpretable explanations themselves provide understanding—qualitative or otherwise. For that, we may have to wait for an explanation to be interpreted and to give rise to another explanation capable of providing the understanding lacking from the first.

It is helpful to put this view of interpretation in vernacular like that used to discuss the general structure of explanation (Figure 5.1). While in explanation the explanandum is some phenomena, diagram, or sentence asserting or describing a phenomenon, in interpretation the thing being interpreted, the *interpretandum*, is an explanation we begin with and that we find difficult to understand. Likewise, while in explanation the explanans is a set of propositions which together explain the explanandum, in an interpretation, the *interpretans* also is an explanation (perhaps conjoined with some additional premises or observations) provided with the intention of being more easily understood. Overall, to provide an interpretation is to show how the interpretans relates to the interpretandum via the *process of interpretation* (Figure 5.4).

Given this general structure of interpretation, we can then begin to classify interpretation strategies and methods. In the remainder of this section, I outline a preliminary classification of interpretation methods with special attention given to their application within MAIS. For instance, the interpretans might be shown by the process of interpretation to be an *approximation* of the interpretandum, it might be *isomorphic* to (but more familiar than) the interpretandum, or it might match the interpretandum only *locally*, within a narrower range of explananda we find more easily understood. These cases are presented in detail below.

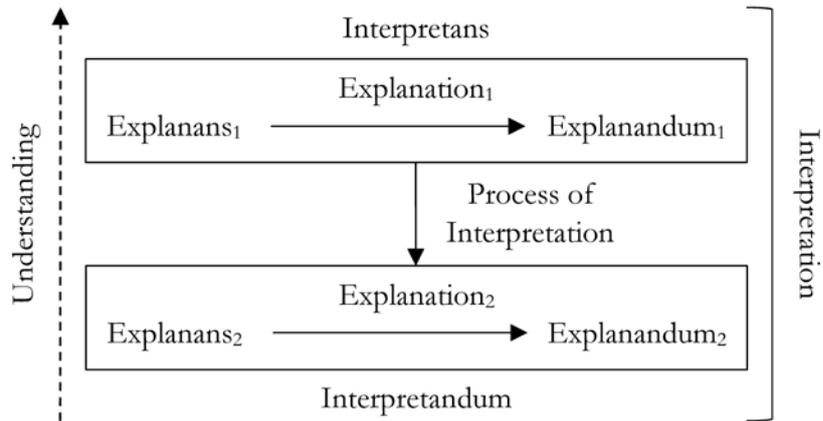


Figure 5.4. The structure of total interpretation.

5.4.2 Total and Partial Interpretation

As noted above, both the interpretans and the interpretandum are explanations; each consists of an explanans, explanation process, and explanandum. A case of *total* interpretation is one in which the interpretans is totally different from the interpretandum. In other words, one in which the explanans, explanation process, and explanandum contained in the former differ in some way from the explanans, explanation process, and explanandum in the latter. For example, consider a simple substitution of variables. Suppose we have an explanation, derived from an ANN in a medical context, that mentions some variable v in each essential part, the explanandum itself being some claim about v , e.g., that it is relevant to breast tissue classification. Coming to this explanation without background information about the meaning of v , we might likewise not understand the meaning of the explanandum. But supposing we are allowed to provide such a meaning by further investigating the set-up of the MAIS, e.g., $v = \text{radiologically white tissue}$, we can construct a new explanation that specifies this meaning in place of v in every essential part, thus changing every part of the explanation. Here, the replacement of “ v ” by “*radiologically white tissue*” in every part of an explanation, thereby generating a new explanation without “ v ” is a process of total interpretation. The new explanation, for example, will not explain that “ v ” is relevant to breast tissue classification, but rather that the presence of radiologically white tissue is relevant to breast tissue classification.³⁸ Often, when we fail to understand an explanation of some phenomenon x , we want to interpret this explanation to provide understanding, but still want to obtain an *explanation of x* . In such cases, what we want is to “adduce” an interpretans, the explanandum of which is identical to the explanandum of the interpretandum (diagrammed in the triangular Figure 5.5, a special case of Figure 5.4). That is, we can sometimes provide a *partial* interpretation by showing how one

³⁸ It should be noted that replacing variables everywhere in an explanation also changes the process of explanation. If, for example, the explanation was DN, the change of variables will change the content of the deduction of the explanandum from the explanans.

explanans arises from another, by some process of interpretation, itself providing some explanation of the very same explanandum. Put another way, a partial interpretation is just a re-explanation of the same explanandum, equipped with a relationship between the new explanans and the old.

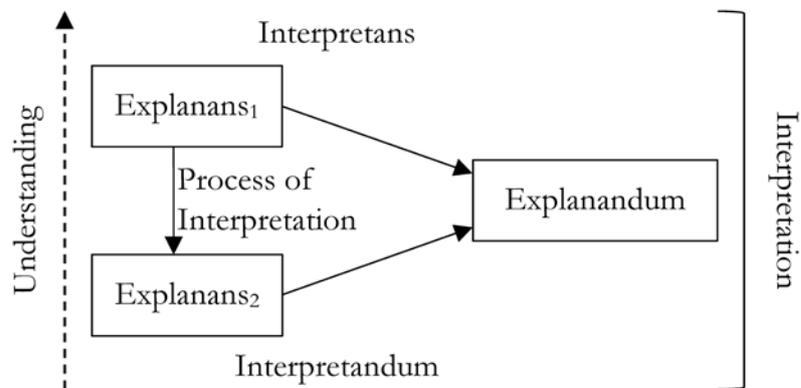


Figure 5.5. A partial interpretation wherein the explanandum remains the same in both the interpretans and interpretandum.

This type of interpretation is often what is aimed for early on in efforts to generate understandable explanations, since the desiderata of the initial explanation of x is understanding of x and remains so after the demand for interpretation. Indeed, another sort of partial interpretation involves interpreting only the process of explanation itself. Suppose we are given a DN explanation where the process of explanation, the *deduction*, is too complicated or long to be held in mind—consider any explanation which requires a mathematical proof that relies on outsourcing steps to a computer, such as Gonthier’s (2005) proof of the four color problem. Here the aim is not to adduce any new explanans or explanandum, these are fixed, but to find some new process of explanation, e.g., a more simple or succinct deduction. This case is diagrammed in Figure 5.6.

Failing a partial interpretation, we may aim to provide a total interpretation. But this immediately raises a problem: what should the relationship between two explanations be such that I come to better understand one by coming to understand the other? If I aim to understand x but its explanation E is unintelligible to me, and I am given a total interpretation such that a second explanation E' explains x' , then there ought to be a special relationship between E and E' if the latter is to provide any understanding whatever about x .

Another way to put the problem of relating different explananda in a total interpretation is by asking what sorts of relationships between phenomena are relevant to scientific explanation: Similar phenomena can sometimes figure in explanations with the same overall structure. For example, explanations of tissue classifications based on radiological whiteness can inform explanations of classifications based on tissue density. Provided we have some justification for substituting radiological whiteness for tissue density, explanations on the basis of radiological whiteness can be totally interpreted by those referring to tissue density instead. The hope being that tissue density provides some understanding that radiological whiteness does not. Of course, the similarity of two

phenomena does not *imply* that there should be any understanding gained about one via explanation of the other—understanding is too psychologically contingent for this—but such relationships are, methodologically, good places to begin the search for understanding.

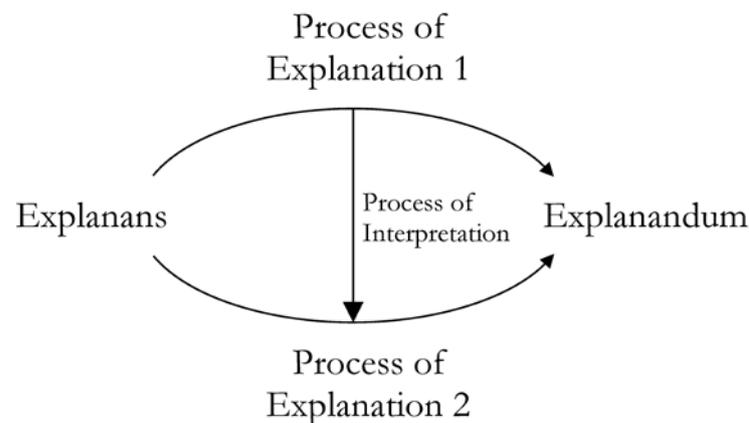


Figure 5.6. A partial interpretation wherein the process of interpretation used results in a new process of explanation for an explanans and explanandum.

In the context of formal explanations, where our concern is not phenomena themselves but the diagrams or descriptions thereof, these observations manifest as strategies or methodologies of interpretation. That is, we can further classify total interpretations according to whether the interpretans and interpretandum are related by approximation or by isomorphism—e.g., if E approximates E' , then we have some reason to think that E' can provide understanding about x . These two cases are detailed below, but since both of these methods come in “local” and “global” forms, we address these notions first.

5.4.3 Local and Global Interpretation

The ML literature is replete with claims about “local” explanation and interpretability (see Ribeiro et al. 2016, 2018; Doshi-Velez and Kim 2017; Adebayo et al. 2018) that, given the accounts of explanation favored among philosophers, are liable to confuse. This is because the accounts of explanation are classified according to the nature of their explanans (laws of nature/statistical laws/entities and activities) and the process of explanation (deduction/induction/depiction), while the designation of an interpretation as “local” is a comparative property of explananda. To our minds the distinction between local and global interpretations is best captured as follows. Considering a pair of explanations E and E' , with respective explananda D and D' , E' is local relative to E if and only if D' is a subset of D . There are many methods of doing this current in ML, however, common to all is the identification of such a subset.

Consider for a common example the Universal Law of Gravitation (*ULG*), which states that the force of gravitation between any two objects is equal to the product of their masses over the square of the distance between them times the gravitational constant. We can use this law in

conjunction with a set of observations (O) about the masses and distances separating each planet in our solar system to explain (or predict) the force that would (absent other interactions) attract every pair of planets (G). Take this to be an explanation $E: ULG + O \rightarrow G$. If we restrict ourselves to explaining only the forces between each planet and earth (L), we may likewise restrict the ULG to the following Earthly Law of Gravitation (ELG): the forces of gravitation between Earth and any other body is equal to the gravitational constant on earth times the mass of the body over the square of the distance of that body from earth. Note that inclusion of explananda can arise due to implication between explanans'. In this case, ULG implies ELG but not vice versa, and this gives rise to an explanation $E': ELG + O \rightarrow L$. Finally, since L is a subset of G , E' is local relative to E .

In the context of ML, a function-based example is also illuminating. Consider the ANN which has learned to output the absolute value of its inputs. A (DN) explanation of the output of this ANN will consist of the statement of the functional “law” together with some input “observations,” these being the explanans, concluding with the statement of the output as explanandum. This can essentially be seen as the function $h(x) = |x|$, giving the explanation $E: “h(x) = |x|” + “O = \{x_i = -1, x_{i+1} = 1, \dots\}” \rightarrow “D = \{x_i = |-1|, x_{i+1} = |1|, \dots\}.”$ Another explanation that only explains the outputs of *positive* inputs could likewise restrict the function $h(x)$ to $x > 0$, which is essentially the same as the function $f(x) = x$, giving the explanation $E': “f(x) = x” + “O = \{x_i = -1, x_{i+1} = 1, \dots\}” \rightarrow “D' = \{x_{i+1} = 1, \dots\}.”$ Since D' is a subset of D , we conclude that E' is local relative to E . Here, to “adduce” a local explanation involved recognizing that $h(x)|_{x>0} = f(x)|_{x>0} = x$, and while in practice applications of constructing local explanations are surely more complex than this, they each share this general theme: once we have decided to explain a sub-set of our explanandum, we can sometimes make this work by adducing a restriction on our explanans—one intended to make the new explanation more understandable. Thus far we have just described “locality” as a comparative property of explanations and based entirely on the inclusion of explananda. In practice, we usually start with one complex explanation, and wish to interpret it, i.e., provide another, more understandable explanation. One way to achieve this is to localize the explanation by adducing another explanation that is local with respect to the explanation we began with. Many methods in ML (e.g., LIME, SHAP) identify this local subset of D by defining a geometric measure of distance between input values and building a model (e.g., a sparse linear model or shallow tree, Section 5.4.4) that approximates the more global model in the environment of a particular input, as defined by this distance metric. Once we have an explanation in hand (the interpretans) given by the locally trained model, we can check to see that indeed it is local (by the process of interpretation) relative to our starting explanation (the interpretandum). The process of interpretation in such cases is showing that the explanans of the interpretans is somehow restricted from the explanans of the interpretandum and demonstrating that the explanandum of the interpretans is included in that of the interpretandum.

In the case of ANNs, our adduction of a restriction on the explanans amounts to a restriction on the inputs to that ANN in essentially the same manner as described for ordinary functional

restrictions. In the most extreme cases of local interpretation we move to an explanation of a single data point, a single classification. And in that somewhat trivializing case, an explanation of the classification of the input can be reduced to a selection of features of the input that most affect the output; such local interpretations provide counterfactual information about which features of the input, if intervened upon, would change the output. For example, local interpretation methods in MAIS often identify the particular pixels, present in an input image, that most affect the classification output (Páez 2019; Zednik 2019; see also Ribeiro et al. 2016). In practice, methods for interpretation of predictions of machine learning algorithms often approximate the explanation, in addition to localizing it to a given datum, as examined in the following section.

5.4.4 *Interpretation by Approximation or Isomorphism*

When an explanation is too complex to be understandable to its user, a possible solution is to interpret the explanation by providing another, similar yet more understandable explanation. In general, measuring similarity depends on a prior choice of some features to compare—mice may be similar to humans with respect to the efficacy of some medication, but not similar in eloquence. In the context of MAIS, an ANN can be thought of as a highly non-linear function $f(x)$. Approximating the ANN predictions would entail providing a (more understandable) function $g(x)$, such that the outputs of $f(x)$ and $g(x)$ are similar on some metric; for example, the square deviation, $\int_x (f(x) - g(x))^2$. When an ANN appears in some explanation, this approximation can successfully figure in the process of interpretation precisely when we expect $g(x)$ to be more understandable than $f(x)$.

Many people find linear functions and decision trees more understandable than ANNs. Indeed, methods for interpretation of ANNs often aim to provide $g(x)$ that is either linear or tree-like, despite what some see as scant justification that these models really are any more understandable (Lipton 2018). Still, some studies illustrate that sparse rule lists and linear models fare well for human-interpretability (Lage et al. 2019; Narayanan et al. 2018). *Distillation* methods work by training one model to approximate the predictions given by another, and can be done either globally or locally.

Linear models are used to provide *local* explanations. While linear models are not flexible enough to globally approximate a neural network, they can provide adequate accuracy when the explanandum is small, such as a single prediction. For this purpose, various methods have been developed that attempt to determine the set of features that are important to explaining the output of a given prediction (Ribeiro et al. 2016; Mishra et al. 2017; Lundberg and Lee 2017; J. Chen et al. 2018). These methods essentially localize the explanation of an ANN prediction to a small neighborhood around a specific prediction (this being the new explanandum) and approximate this localized explanation. For example, for the ANN described above for the prediction of dense breast tissue, while a global explanation of what makes tissue dense might be too complex to understand, an explanation of approximately which pixels in a given mammogram led to it being classified as dense may be simple enough to be understood by visual inspection.

Since linear models are often not flexible enough to provide sufficient global approximations, global ANN distillation in particular involves using decision trees as the approximating model. When trees are more understandable, distillations are thereby a method of interpretation by approximation. Trees and ANNs both have quite a strong property *as approximators*. Both are *universal approximators*, meaning that they can approximate any continuous function to an arbitrary degree of accuracy (Hornik et al. 1989). Since ANNs are continuous functions, decision trees can therefore approximate neural networks to any accuracy.

The process of distilling a neural network into a tree involves generating input data and using ANNs to get a prediction, then training a decision tree on the sets of input-output pairs given by the ANN. One can always generate as much data as needed to achieve a desired level of similarity between the model outputs. Nonetheless, despite the *prima facie* understandability of *small* trees as compared to networks, in practice the distillation of useful ANNs tends to result in *very large* trees (Frosst and Hinton 2017).

Here we are faced with the understandability-accuracy trade-off, since one can also sacrifice accuracy of a tree distillation by reducing or fixing its size or depth (Zhou et al. 2018), thereby presumably purchasing some understandability at the expense of accuracy. Though we may gain some traction of understanding by initially approximating an ANN, the upshot of this is that, as we increase the accuracy of the approximating model, we wash out any understanding gained by approximation itself and are left only with whatever advantage is offered by modelling (nearly) the same process in a different sort of model. When we interpret some explanans that includes an ANN into another including a tree, which approximates the first to an arbitrary degree, we have, in a way, gone beyond approximation into interpretation by isomorphism.

Since understanding has a psychological component, it has some mathematically not well-behaved properties. For one, understanding is not isomorphism invariant. That is, if I understand a , the fact that a is isomorphic to b does *not* imply that I understand b . Intuitively, my understanding of some statement of English does not imply anything about my understanding of equivalent sentences in another language. This is because our understanding can turn on features of the specific presentation of some notion and the background knowledge of the user of the explanation. Nonetheless, the converse of this has been put to use as a method of interpretation. Indeed, if I do not understand a , I may still understand some isomorphic b (due perhaps to my familiarity with constructions in the language of b).

Perhaps the ubiquity of trees and tree-like structures in our everyday experience explains the prevalence of tree distillations in ML; their familiarity evidently leads to the idea that they will be understandable, and that may indeed be a reasonable assumption to make. Nonetheless, there is a very wide range of types of models that can universally approximate, or are genuinely isomorphic to, ANNs. Candidate alternative universal approximators include fuzzy systems (Zadeh 1965; Wang 1992; Kosko 1994; Castro 1995; Yen et al., 1998, cf. Klement et al. 1999); Neural ODEs (R. Chen et al. 2018; Zhang et al. 2019; and references therein) and nearest-neighbor methods (Cover and Hart

1967; Stone 1977). When we have reason to believe that any of these alternative models are “more understandable” than ANNs, we thereby have reason to make use of these isomorphic models to provide an interpretans, i.e., an explanation using the alternative model within its explanans, to serve the same role that the ANN served in the interpretandum.

5.5 Interpretation and Understanding

Up until this point, I have made the case for the explainability of ANNs and MAIS and argued that ML researchers have resources for interpretation which allow them to produce more understandable explanations of AI systems’ outputs. It may be argued that what is really at stake in discussions about the opaque nature of ML models is not about complexity, but rather about the human capacity to understand, in a particular way, *why* certain outputs were generated. To clarify, an ANN is trained on some set of data, and it discovers some patterns and associations in the data that are important for making accurate recommendations. While the discovered associations and patterns may be articulated in terms of weightings in the ANN, no human can really understand the reasons why those weightings are generated by the model. In contrast, other similarly complex machinery, such as rockets that carry astronauts into orbit or early computerized chess engines, were developed using human knowledge. Because these complex machines were developed with existing human knowledge, there are at least some humans who can understand them. Where ML models, in a sense, make decisions based on their own discoveries, other complex machines are developed using human knowledge. So, ML models discover features that are useful in making decisions, but, because these features were not the result of human discovery or based on human knowledge, the kind of understanding we gain from the kinds of explanations generated by interpretations is impossible to evaluate. It may be inferred that the account of interpretation provided here does little to solve this problem.

The above concerns may be unpacked with the distinction between *objectual understanding* and *explanatory understanding* (Kvanvig 2009). Objectual understanding can be characterized in the following way: One objectually understands an explanation if they grasp that its explanans entails or is probabilistically or causally relevant to its explanandum. Put another way, the person has *that*-understanding. It seems that interpretations, as they are presented here, provide this sort of understanding—it makes it clear to the subject *that* the original explanation’s explanans bears some relation to its explanandum, be it deductive, probabilistic, or causal. Take our running example, the MAIS which classifies breast tissue. By approximating which values in the input layer of the neural network led to tissue being classified as dense, all we seem to be doing is illustrating that these values are probabilistically relevant to the tissue being dense. Understanding that there is a probabilistic relation between the values and the classification, it may be argued, is not useful for evaluating the MAIS’ output.

However, what we want to understand is *why* the tissue was classified as dense. Not only would this allow us to assess the trustworthiness of the output, but also be able to provide patients with some sort of narrative regarding why the diagnosis was generated by the MAIS. Of course, it is not

the case that MAIS decisions and recommendations have sole jurisdiction over diagnoses, prognoses, and predictions in the clinic, but since they factor into such judgements, explaining the reasons why their outputs should be taken seriously is crucial. This sort of understanding is explanatory understanding: One explanatorily understands an explanation occurs if they grasp why the explanans entails or is probabilistically or causally relevant to its explanandum (see Kvanvig 2009). Put another way, the person has *why*-understanding. Interpretations of ANNs, as they are presented here, do not seem to provide this type of understanding.

It has been argued, however, that objectual understanding can be reduced to explanatory understanding. Here, Khalifa (2017) suggests that particular cases of objectual understanding can be matched to a corresponding explanatory understanding. Correspondingly, we may be able to reduce the objectual understanding provided by AI-interpretation to the explanatory understanding we seek. For example (abstracting much detail), the interpretation of the tissue classifier MAIS above approximates the input layer values which led to the output classification. These values, however, correspond to pixels in the mammogram image in question. We can thus translate the values such that we the parts of the mammogram image which led to the categorization can be identified. Indeed, just these kinds of strategies, such as saliency methods (Adebayo et al. 2018), are being increasingly investigated in ML research.

A more thorough analysis of the reduction of objectual understanding provided by interpretations of ANNs to explanatory understanding is beyond the scope of this Chapter, but the upshot is this: After an interpretation is performed, it may be possible to obtain explanatory understanding from the objectual understanding provided by the interpretation. However, it should be noted that this relies on having objectual understanding in the first place, which in turn relies on successfully interpreting ANNs.

5.6 Conclusion

Conceiving of explanation according to those accounts offered within the philosophy of science disentangles the accuracy-explainability trade-off problem in AI, and in doing so, deflates the apparently paradoxical relationship between trust and accuracy which seems to plague debates in medical AI and the ML literature. If it is simply that explainability is required for trust, there is no cause for worry, since highly accurate (and potentially complex) MAIS are just as explainable as simple ones. I resolve this issue by using clearly defined accounts of explanation, and demarcating notions of explanation from understanding and interpretation. The account of interpretation provided here is explicitly positioned as distinct from accounts of explanation, while making clear the connection that both have to understanding.

While explanation is a well-theorized notion within the theory and philosophy of science, “interpretation” and a corresponding notion of “interpretability” are not (see Lipton 2018). I have attempted to synthesize a notion of scientific “interpretability,” from cases where it was *not* used synonymously with either “explainability” or “understandability.” This was done to provide a theoretical framework that generalizes the methods scientists and ML researchers often use for the

purpose of interpretation, and to help remedy this lacuna within the ML literature in particular and the neglect of inter-explanatory relationships in philosophy of explanation broadly speaking. In the context of MAIS, there never really was a problem explaining artificial networks. Rather, the problem has always been understanding the explanations that were available, and our account of interpretation shows why.

Interpretation is a process taking one explanation to another, more understandable explanation. Understandability is at least partly psychological, depending not only on the phenomenon or its explanation but also on the user of the explanation. Identifying the elements of explanations that make them understandable broadly speaking is an active research topic in the social sciences (Miller 2019), but beyond the scope of a theory of interpretation as such. Just as de Regt (2017) says we need a general theory of understanding that is independent of our specific accounts of explanation, we require a general theory of interpretation that is independent of both. The framework I provide can thereby only be employed once we have made some theoretical choices about what features of explanations indeed provide understanding, without which the success of an interpretation cannot be assessed.

The ubiquitous presumption that only simple and/or linear models are “understandable” is liable to limit the potential scope of scientific interpretation; the use of non-linear and complex models should not be excluded at the outset. It seems a large part of the stress on the point of explainability in discussions of ANNs and MAIS boils down to an insistence that they be understandable to a non-specific and correspondingly broad audience of clinicians, patients, and possibly the general public. With such a diverse audience of users of explanations, perhaps simplicity is the only proxy for understandability, and persistent demands for “explainable” ANNs are reducible to demands for simple and potentially correspondingly weak MAIS. We can move away from this tendency to simplicity by demanding that ANNs be *interpretable* in the sense defined here. That is, by demanding that we find ways to convert explanations that are not understood into those that are more understandable in a user-relative way. That way, we might keep many complex and thus strong MAIS while achieving broad understandability by, counter-intuitively, adding further “simplifying complexity” in the form of interpretation methods.

Conclusion

This dissertation has explored several conceptual and methodological features related to predicting the effectiveness of medical interventions. While I do not defend any explicit overarching thesis, the arguments presented here are primarily motivated by concerns over how conclusions about the general capacity of medical interventions can be improved. There is much room for progress in this regard. As I have illustrated throughout the dissertation, important concepts can be conflated, clinical methods are commonly exploited, significant forms of evidence are often left out of therapeutic predictions, and emphasis is placed on ineffective strategies for improving medical inferences. In this conclusion, I give a brief recap of each chapter and discuss two broad themes that have emerged from the project.

Here is a summation of the key topics and claims from each chapter. In the Introduction, I provided a definition of medical effectiveness to work with throughout the rest of the dissertation. This definition was based on philosophical insights from Cartwright (2009; 2012) and Stegenga (2018). I characterized medical effectiveness as a claim about an intervention's capacity to prevent or modulate the physiological dysfunction or disvalued state of a disease generally or in a target context.

In Chapter 1, "A Pragmatic Approach to Disease Classification", I clarified three prominent models of disease classification: the etiological, symptom-based, and pathophysiological models. I argued for a pragmatic approach to disease classification whereby one's choice of model should be based on what is best suited for their aims, suggesting that the pathophysiological model should be most appropriate for the goal of predicting medical effectiveness.

In Chapter 2, titled "Integrating Mechanistic Evidence for Inferring Medical Effectiveness", I provided an informal approach for the systematic integration of mechanistic evidence with statistical evidence.

Chapter 3, "P-hacking: Its Costs and When It Is Warranted" focused on the questionable research practice of p-hacking. I used philosophical tools from decision theory to articulate the prominent view that p-hacking is both epistemically and practically harmful. Using this framework allowed me to illustrate flaws in the standard argument against p-hacking. I then argued that there are scenarios in which practices which amount to p-hacking may be warranted.

In Chapter 4, "Curb Your Effectiveness: Correcting for Meta-Biases in Therapeutic Prediction", I argued estimations of medical effectiveness are liable to fail because of a particularly harmful set of biases, which I refer to as meta-biases. Such biases lead to systemic overestimations of medical effectiveness. I proposed the bias dynamics model as a solution to the problem of meta-biases, arguing for the use of bias coefficients to correct estimated of medical effectiveness based on meta-research evidence about meta-biases.

Finally, in Chapter 5, titled “Medical Artificial Intelligence: What is Interpretability?”, I proposed a novel theory of AI-interpretability to help remedy confusion in discussions of AI models. The confusion, I argued, is due to widespread equivocation on the concepts of explainability, understandability, and interpretability in the literature. The theory of interpretability I proposed can help clarify the important relationships between these concepts.

Three broad themes have emerged during the course of this project: the role of mechanisms and mechanistic evidence in medical inference; the importance of pragmatic considerations in clinical research methodology; and the links between values and epistemic concerns in medicine. These themes present promising avenues for future research in medical epistemology generally, and predicting medical effectiveness in particular.

Starting with the least novel of these, I have at multiple points in the dissertation discussed the role of *mechanisms* and *mechanistic evidence* in medical inference. Take the arguments presented in Chapter 1 on disease classification. I argued that pathophysiological classifications of disease are often instrumental to predicting whether a particular treatment will be efficacious. This is because pathophysiological classifications give us information about the underlying mechanisms of a disease, and predicting medical effectiveness sometimes relies on understanding the underlying mechanisms of a given disease. In Chapter 2, I do some clarificatory work on the notion of mechanistic evidence as it appears in discussions about evidence and medical inference. And in Chapter 5, mechanisms play a central role in two models of explanation that may be used to provide interpretations of artificial neural networks.

Since the first articulation of new mechanist philosophy by Bechtel and Richardson (1993), mechanisms have become a dominant feature of philosophy of science, especially in philosophy of biology and philosophy of medicine (Machamer et al. 2000; Glennan 2002; Bechtel and Abrahamsen 2005; Steel 2008; Levy 2013; Parkkinen et al. 2018; Marchionni and Reijula 2019). Despite its ubiquity in discussion about medical evidence, there is more work to be done. For instance, given the many calls for including mechanistic evidence in medical inference (e.g., Parkkinen et al. 2018), further exploration of techniques for amalgamating mechanistic evidence with other types of evidence is needed, much like that of Landes et al. (2018), De Pretis et al. (2019), and Park et al. (*forthcoming*). Such techniques, while promising, are still nascent in their development and applications. Moreover, there are extensive literatures on robustness in science (e.g., Weisberg 2006; Woodward 2006; Parkkinen 2016; Stegenga and Menon 2017) and evidence amalgamation (e.g., Stegenga 2013; Fletcher et al 2019; Holman 2019) which is relevant to debates about the role of mechanistic evidence in medicine.

Another important set of concerns involves the question of whether mechanisms constitute a worthwhile ontological basis for medical epistemology. For example, one might argue that a *process ontology*, like that proposed by Whitehead (1929) and, more recently, Dupré (2020) could be helpful in medical science. Process ontology has primarily been proposed in philosophy of biology. On this view, our understanding of biology has been hindered by wrongly assuming that the world

is fundamentally composed of entities and their activities, i.e., mechanisms. Rather than adopting this such a substance ontology, the argument goes, we should think of the world as fundamentally composed of *processes*, i.e., just activities. What we think of as fixed entities are in fact momentary stabilities in a flux of processes. Given the reliance on mechanistic thinking in biomedical research, it stands to reason that the arguments for process ontology in biology apply to medicine too. Dupré for example, argues that germ theory, which is based on mechanistic thinking, has led to categorization microbes as “good, bad, or indifferent” (2020: 108). This, he argues, is untenable since there are cases such as bacteria that are necessary for a healthy digestive system causing critical disease in other parts of one’s body. Processual thinking, for Dupré, can help by moving us away from seeing microbes as entities with intrinsic properties. There is more room for research on how mechanistic thinking and processual thinking stack up against one another in medicine.

A second theme that has emerged in this dissertation is the role of *pragmatic considerations* in medical research. Most of the arguments made here are based on what works best for the goal of predicting medical effectiveness, or what is useful for medical inference, or understanding. A stark example of this comes in Chapter 1, where I argue for a pragmatic approach to disease classification. However, I gesture toward this sort of thinking in the other chapters of the dissertation too. In Chapter 2, for instance, I suggest that one important practical consideration when deciding between using a mathematical evidence amalgamation technique or a systematic evidence amalgamation technique is whether we have access to either quantitative or qualitative information about the relevant features of the study and target populations in question. What is more, different stakeholders often have varied capacity to understand how we arrive at specific conclusions. Policy decision makers, for example, may not be as well-versed in the quantitative techniques of medical research, and many individuals in the public even less so. The explanations we provide for the ways in which we infer conclusions using varied evidence should be catered according to audience. This pragmatic consideration features in Chapter 5 as well. Interpreting artificial neural network outcomes involves deriving an understandable explanation from one we find difficult to understand. Interpretability is geared toward pragmatics about explaining artificial intelligence models. And in Chapter 3, I argue that we should consider the practicalities of a situation when deciding whether p-hacking is warranted. The same can be said of other questionable research practices, such as hypothesizing after results are known and data fishing. Finally, in Chapter 4, there are pragmatic elements to proposing that the bias dynamics model make use of evidence about meta-biases, not methodological biases. Methodological biases are typically difficult to detect and when we do find evidence for them, there is often disagreement about their effects on the quality of evidence.

Most, if not all, the features related to predicting medical effectiveness, and indeed most areas of medical epistemology, can be linked to pragmatic factors. Medicine is, after all, a practical endeavor. Given this, one might argue that pragmatic considerations are a necessary element for successful medical inference. There is an opportunity to develop an entire research program on the topic of *medical pragmatism*. Such a project would unpack the relevant pragmatic features of

medical reasoning, outlining tensions between each, and provide an account of how to navigate these various features in pursuit of reliable medical inference. One upshot of medical pragmatism, so construed, is that medical inference should be articulated as a multifaceted process involving, among other things deeply pragmatic considerations, and not, as it is often characterized by proponents of EBM, something that follows a strict set of principles.

The last theme to emerge from the discussions in this dissertation, and one that is inextricably tied to pragmatic considerations, involves *the intersection of values and epistemic concerns* in medicine, particularly with the use of medical AI systems. One explicit example of this comes in Chapter 3, where I appeal to the argument from inductive risk to highlight the role of value judgments in warranting the practice of p-hacking. I argue that p-hacking can be warranted in scenarios where there is urgent need for discovery, such as the onset of a global pandemic. Naturally, given that a primary aim of medicine is to prevent death and mitigate suffering, medical inference should be fraught with value-based judgements. The argument from inductive risk can be applied to claims over whether mechanistic evidence should be considered for causal inference in medicine (see Chapter 2). Abstracting much detail, one might argue that if the consequences of wrongly deploying an intervention were possibly severe, then we should aim to include available mechanistic evidence. For instance, given that the consequences of wrongly deploying efavirenz in the Zimbabwean context could have been (and in fact were) serious, the mechanistic evidence, which those involved knew existed, should have been considered. Of course, debates over the role of values in science are not new (cf. Rudner 1953; Douglas 2009; Betz 2013). However, in Chapter 5, the debates about explainability, understandability, and interpretability in AI are partially predicated on ethical and value-based considerations. Often, the reason we want to interpret an artificial neural network is because it can help explain the ethical concerns of medical AI in understandable ways, with the hope that we can resolve the issues.

One such set of concerns centers on instances where using medical AI systems has led to unfairness and discrimination in clinical contexts. An example of this is an algorithm that is less likely to allocate resources to black patients than white patients who are equally sick (Rajkomar et al. 2018). Given such cases, several AI ethicists have argued that machine learning models, especially those intended for use in healthcare, should be designed with equitable outcomes in mind (Mittelstadt et al. 2016; Morley and Floridi 2020). This often requires that we be able to understand the AI models we use, something I discuss in the dissertation. There is more to be said about how understanding, related to ethical concerns, particularly in relation to medical equity.

One important set of questions involve the kind of understanding we can acquire from AI-interpretations. One potential problem here is that current interpretation processes seem to only provide *objectual understanding* of ML models. One objectually understands an explanation if they grasp that (part of) its explanans entails or is probabilistically or causally relevant to its explanandum (Kvanvig 2009). In other words, an objectually understandable explanation tells us *that* some numerical values in the model are relevant to the output or recommendation provided.

Approximating the predictions of a health resource allocation algorithm may help us understand that the numerical values connected to one's income level are relevant to a particular resource allocation score. However, what we typically want to understand is *why* a patient received a certain resource allocation score. This is *explanatory understanding*: One explanatorily understands an explanation if they grasp why the explanans entails or is probabilistically or causally relevant to its explanandum (ibid.). Interpretations of algorithms may not be able to provide this type of understanding. More research is required on whether objectual understanding or explanatory understanding is better for preventing potential unfairness from using medical AI. Developing a deeper understanding of the links between the concepts of understanding and equity in AI is crucial to finding novel solutions to the challenges facing the evaluation of medical algorithms and mitigating unfair healthcare outcomes that may result from their use.

These concluding remarks point to the need for more philosophical work on predicting the effectiveness of medical interventions. That being said, it is my hope that this dissertation has helped identify some of the key concerns underlying the roles of evidence, bias, and artificial intelligence in therapeutic prediction, and provided insight into how to think about them.

Appendix 1: The Expected Utility Principle

The arguments I develop in Chapter 3 are based on the *expected utility principle* (EUP). The standard EUP states that an act a_1 is preferable to another act a_2 if and only if the expected utility of a_1 is greater than the expected utility of a_2 . This is represented formally as follows:

$$a_1 \succ a_2 \text{ iff } EU(a_1) > EU(a_2)$$

This principle is best illustrated by example. Suppose Mandy has decided to get some writing done. The coffee shop she normally works at can sometimes be full which means there would not be a table for her, in which case she would have wasted valuable time in the trip there. She could stay at home and get an average amount of work done, but she would rather sit at the atmospheric coffee shop since it increases her productivity. Mandy is faced with a decision between going to the coffee shop or staying at home to write.

	Space at coffee shop	Coffee shop is full
Go to the coffee shop	Increased productivity	Lose valuable time
Stay home	Average productivity	Average productivity

Table A

Mandy's decision can be formulated in terms of three kinds of entity: *states*, *acts*, and *outcomes* depicted in Table A. Let $S = \{s_1, s_2, \dots, s_n\}$ be the set of all states of the world out of the decision maker's control. In the example, there are two states relevant to Mandy's decision, there is *space at the coffee shop* (s_1), or the *coffee shop is full* (s_2). Let $A = \{a_1, a_2, \dots, a_n\}$ be the set of acts the decision maker can choose from. For simplicity, in the example Mandy has two relevant choices, either *go to the coffee shop* (a_1) or *stay home* (a_2). Outcomes are the consequences of a decision maker's different acts that occur under different states of the world. For Mandy, there are three possible outcomes: *Increased productivity* (o_1) from her going to the coffee shop when there is space for her to work; *losing valuable writing time* (o_2) should she go to the coffee shop and find out it is full; and *average productivity* from her staying at home (the coffee shop being full or empty has no effect on her productivity here).

According to EUP, making a decision here requires calculating the expected utility each act. The expected utility of an act a_i is equal to the sum of the products of the utility of the outcome of the act, $U(o_j)$, and the probability of the state occurring $P(s_k)$:

$$EU(a_i) = \sum_{k=1}^n U(o_j) P(s_k) \quad (\text{EU1})$$

Applied to Mandy's decision problem, the expected utility of going to the coffee shop would be the product of the utility of having increased productivity and the probability of there being space at the coffee shop, added to the product of the utility of losing valuable writing time and the

probability that the coffee shop is full. EU1 thus provides a weighted average of the values of the states given the decision maker's acts, where the weights are the probabilities of the states. Using EUP, one can act according to how valuable the outcomes are to the decision maker. If the expected utility of going to the coffee shop is greater than that of staying home to work, then according to EUP, Mandy should go to the coffee shop.

One problem with this definition of expected utility is that it ignores any connection that may exist between an act and a state. In other words, EU1 does not make room for cases where an act increases the probability of one possible state over others. Take the following example: Say that Mandy has now arrived at the coffee shop and found that there is space for her to work. She has a few projects on the go but knows that she is facing a deadline for a specific paper. Mandy could either work on this paper at while she is at the coffee shop or work on another project (i.e., not work on this paper). She reasons that she could work on that paper another time and possibly still meet her deadline and either she will meet her deadline, or she will not meet her deadline. Mandy faces the decision problem outlined in Table B. Given that the probability of her meeting the deadline is increased by working on her paper (and likewise the probability of her not meeting the deadline is increased by not working on the paper), the expected utilities of the acts would be affected by her choice. *EU1* does not account for such cases.

	Meet deadline	Fail to meet deadline
Work on paper	o_1	o_2
Do not work on paper	o_3	o_4

Table B

Given such instances, the probability of the state is usually interpreted as a conditional probability. This allows a rational decision maker to choose the act which if performed will most likely ensure that she achieve her desired outcomes. Such an interpretation replaces the probability of the state $P(s_k)$ in EU1 with the conditional probability $P(s_k|a_i)$:

$$EU(a_i) = \sum_{k=1}^n U(o_j) P(s_k|a_i) \tag{EU2}$$

Here, $P(s_k|a_i)$ is the probability of state s_k given act a_i and is defined as $\frac{P(s_k \cap a_i)}{P(a_i)}$. It measures the degree of confidence the decision maker would have in state s_k occurring if a_i were chosen. For Mandy, her working on the paper in question increases her confidence that she will meet her deadline, hence EU2 provides Mandy with a way to order her preferences over the choice to work on her paper or not. In Chapter 3, I use EU2 in my application of the expected utility principle.

Appendix 2: Outcome Measures

In clinical trials, measures of efficacy are quantified as effect sizes using many available outcome measures. Outcomes can be continuous (such as total serum cholesterol level) or dichotomous (such as healthy cholesterol levels or unhealthy cholesterol levels). In this dissertation, for simplicity, I have focused on dichotomous outcomes. Dichotomous outcomes can be measured using either absolute or relative outcome measures. I describe the most prominent of these outcome measures below.

Consider a trial that assesses whether the exposure to some intervention is correlated with a particular dichotomous outcome. The study set up is described in Table C, where O_E , O_C , N_E , and N_C with outcome O and no outcome N in the group receiving the treatment, (E for ‘experimental group’) and the group receiving either placebo or no treatment, C (for ‘control group’). T_E , and T_C are the total numbers of participants in the experimental and control groups respectively.

	O	N	T
Experimental Group	O_E	N_E	T_E
Control Group	O_C	N_C	T_C

Table C

Commonly used relative outcome measures for dichotomous outcomes are relative risk (RR), relative risk reduction (RRR), and odds ratio (OR). These are defined as follows:

$$RR = \left(\frac{O_E}{T_E}\right) / \left(\frac{O_C}{T_C}\right)$$

$$RRR = \left[\left(\frac{O_E}{T_E}\right) - \left(\frac{O_C}{T_C}\right) \right] / \left(\frac{O_C}{T_C}\right)$$

$$OR = \left(\frac{O_E}{N_E}\right) / \left(\frac{O_C}{N_C}\right)$$

Commonly used absolute outcome measures for dichotomous outcomes include absolute risk reduction (ARR) and number needed to treat (NNT), defined as follows:

$$ARR = \left(\frac{O_E}{T_E}\right) - \left(\frac{O_C}{T_C}\right)$$

$$NNT = 1 / \left[\left(\frac{O_E}{T_E}\right) - \left(\frac{O_C}{T_C}\right) \right]$$

References

- Achinstein, P. (1983) *The nature of explanation*. New York: Oxford Univ. Pr.
- Adebayo, J. *et al.* (2018) ‘Sanity checks for saliency maps’, 31, pp. 9505–9515.
- Adkins, D. E. (2017) ‘Machine learning and electronic health records: a paradigm shift’, *American Journal of Psychiatry*, 174(2), pp. 93–94.
- Agustí, A. *et al.* (2017) ‘Precision medicine in airway diseases: moving to clinical practice’, *European Respiratory Journal*, 50(4), p. 1701655.
- Aklillu, E. *et al.* (2007) ‘Pharmacogenetics of cytochrome P450s in African populations: clinical and molecular evolutionary implications’, in Suarez-Kurtz, G. (ed.) *Pharmacogenomics in Admixed Populations*. Austin, TX.: Landes Bioscience, pp. 99–119.
- Albarqouni, L. N., López-López, J. A. and Higgins, J. P. T. (2017) ‘Indirect evidence of reporting biases was found in a survey of medical research studies’, *Journal of Clinical Epidemiology*, 83, pp. 57–64.
- Aler Tubella, A. *et al.* (2019) ‘Governance by glass-box: Implementing transparent moral bounds for AI behaviour’, in *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence. Twenty-Eighth International Joint Conference on Artificial Intelligence {IJCAI-19}*, Macao, China: International Joint Conferences on Artificial Intelligence Organization, pp. 5787–5793.
- American Psychiatric Association (1980) *Diagnostic and statistical manual of mental disorders: DSM-III*. 3rd ed. Washington, DC: American Psychiatric Association.
- American Psychiatric Association (1987) *Diagnostic and statistical manual of mental disorders: DSM-III-R*. 3rd ed. revision. Washington, DC: American Psychiatric Association.
- American Psychiatric Association (1994) *Diagnostic and statistical manual of mental disorders: DSM-IV*. 4th ed. Washington, DC: American Psychiatric Association.
- American Psychiatric Association (2000) *Diagnostic and statistical manual of mental disorders: DSM-IV-TR*. 4th ed., text revision. Washington, DC: American Psychiatric Association.
- American Psychiatric Association (2013) *Diagnostic and statistical manual of mental disorders: DSM-5*. 5th ed. Washington, D.C: American Psychiatric Association.
- Andersen, H. (2012) ‘Mechanisms: what are they evidence for in evidence-based medicine?: Mechanisms evidence for what?’, *Journal of Evaluation in Clinical Practice*, 18(5), pp. 992–999.
- Andersen, H. (2014a) ‘A field guide to mechanisms: Part I’, *Philosophy Compass*, 9(4), pp. 274–283.
- Andersen, H. (2014b) ‘A field guide to mechanisms: Part II’, *Philosophy Compass*, 9(4), pp. 284–293.
- Angell, M. (2005) *The truth about the drug companies: how they deceive us and what to do about it*. Rev. and updated. New York: Random House Trade Paperbacks.

- Anglemyer, A., Horvath, H. T. and Bero, L. (2014) 'Healthcare outcomes assessed with observational study designs compared with those assessed in randomized trials', *Cochrane Database of Systematic Reviews*. Edited by Cochrane Methodology Review Group, 2014(4).
- Antman, E. M. and Loscalzo, J. (2015) 'Ischemic heart disease', in Kasper, D. et al. (eds) *Harrison's Principles of Internal Medicine*. 19th Edition. New York, NY: McGraw-Hill Education.
- Anvari, F. and Lakens, D. (2018) 'The replicability crisis and public trust in psychological science', *Comprehensive Results in Social Psychology*, 3(3), pp. 266–286.
- Apostolova, N. et al. (2015) 'Efavirenz and the CNS: what we already know and questions that need to be answered', *Journal of Antimicrobial Chemotherapy*, 70(10), pp. 2693–2708.
- Armstrong, D. (2011) 'Diagnosis and nosology in primary care', *Social Science & Medicine*, 73(6), pp. 801–807.
- Athey, S. (2017) 'Beyond prediction: using big data for policy problems', *Science*, 355(6324), pp. 483–485.
- Banks, G. C. et al. (2016) 'Questions about questionable research practices in the field of management: a guest commentary', *Journal of Management*, 42(1), pp. 5–20.
- Bechtel, W. (2006) *Discovering cell mechanisms: The creation of modern cell biology*. New York: Cambridge University Press.
- Bechtel, W. (2011) 'Mechanism and biological explanation', *Philosophy of Science*, 78(4), pp. 533–557.
- Bechtel, W. and Abrahamsen, A. (2005) 'Explanation: a mechanist alternative', *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, 36(2), pp. 421–441.
- Bechtel, W. and Richardson, R. C. (1993) *Discovering complexity: Decomposition and localization as strategies in scientific research*. Princeton: Princeton University Press.
- Bechtel, W. and Richardson, R. C. (2010) *Discovering complexity: Decomposition and localization as strategies in scientific research*. MIT Press ed. Cambridge, Mass: MIT Press.
- Bekelman, J. E., Li, Y. and Gross, C. P. (2003) 'Scope and impact of financial conflicts of interest in biomedical research: a systematic review', *JAMA*, 289(4), pp. 454–465.
- Bero, L. et al. (2007) 'Factors associated with findings of published Trials of drug–drug comparisons: why some statins appear more efficacious than others', *PLoS Medicine*. Edited by A. Liberati, 4(6), p. e184.
- Berwick, D. M. (2005) 'Broadening the view of evidence-based medicine', *Quality and Safety in Health Care*, 14(5), pp. 315–316.
- Betz, G. (2013) 'In defence of the value free ideal', *European Journal for Philosophy of Science*, 3(2), pp. 207–220.
- Bhandari, M., Bhandari, Aakanksha and Bhandari, Anil (2011) 'Recent updates on codeine', *Pharmaceutical Methods*, 2(1), pp. 3–8.

- Biddle, J. (2007) 'Lessons from the Vioxx debacle: what the privatization of science can teach us about social epistemology', *Social Epistemology*, 21(1), pp. 21-39.
- Bird, A. (2020) 'Understanding the replication crisis as a base rate fallacy', *The British Journal for the Philosophy of Science*.
- Bolton, D. (2012) 'Classification and causal mechanisms: A deflationary approach to the classification problem', in Kendler, K. S. and Parnas, J. (eds) *Philosophical Issues in Psychiatry II*. Oxford: Oxford University Press, pp. 6–11.
- Boorse, C. (1977) 'Health as a theoretical concept', *Philosophy of Science*, 44(4), pp. 542–573.
- Broadbent, A. (2009) 'Causation and models of disease in epidemiology', *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, 40(4), pp. 302–311.
- Broadbent, A. (2013) *Philosophy of epidemiology*. New York: Palgrave Macmillan (New directions in the philosophy of science).
- Bruns, S. B. and Ioannidis, J. P. A. (2016) 'p-curve and p-hacking in observational research', *PLOS ONE*. Edited by D. Marinazzo, 11(2), p. e0149144.
- Buckner, C. (2019) 'Deep learning: A philosophical introduction', *Philosophy Compass*, 14(10).
- Campaner, R. (2011) 'Understanding mechanisms in the health sciences', *Theoretical Medicine and Bioethics*, 32(1), pp. 5–17.
- Campbell, D. T. (1957) 'Factors relevant to the validity of experiments in social settings', *Psychological Bulletin*, 54(4), pp. 297–312.
- Carter, K. C. (2003) *The rise of causal concepts of disease: Case histories*. Burlington, VT: Ashgate.
- Cartwright, N. (2007) 'Are rcts the gold standard?', *Biosocieties*, 1, pp. 11–20.
- Cartwright, N. (2009) 'What is this thing called "efficacy"?', in Mantzavinos, C. (ed.) *Philosophy of the social sciences : philosophical theory and scientific practice*. Cambridge: Cambridge University Press, pp. 185–206.
- Cartwright, N. (2012) 'Presidential address: Will this policy work for you? Predicting effectiveness better: How philosophy helps', *Philosophy of Science*, 79(5), pp. 973–989.
- Caruana, R. *et al.* (2015) 'Intelligible models for healthcare: predicting pneumonia risk and hospital 30-day readmission', in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '15. the 21th ACM SIGKDD International Conference*, Sydney, NSW, Australia: ACM Press, pp. 1721–1730.
- Castro, J. L. (1995) 'Fuzzy logic controllers are universal approximators', *IEEE Transactions on Systems, Man, and Cybernetics*, 25(4), pp. 629–635.
- Centers for Disease Control and Prevention (2019) *CDC - global health - noddling syndrome*. Available at: <https://www.cdc.gov/globalhealth/noddingsyndrome/default.htm> (Accessed: 30 April 2021).

- Chatterjee, S. *et al.* (2015) ‘Draft genome of a commonly misdiagnosed multidrug resistant pathogen *Candida auris*’, *BMC Genomics*, 16(1), p. 686.
- Chen, J. *et al.* (2018) ‘Learning to explain: An information-theoretic perspective on model interpretation’, *arXiv:1802.07814 [cs, stat]*.
- Chen, R. T. Q. *et al.* (2018) ‘Neural ordinary differential equations’, *arXiv:1806.07366 [cs, stat]*.
- Chou, R. *et al.* (2016) *Statin use for the prevention of cardiovascular disease in adults: a systematic review for the U.S. preventive services task force*. Rockville (MD): Agency for Healthcare Research and Quality (US) (U.S. Preventive Services Task Force Evidence Syntheses, formerly Systematic Evidence Reviews).
- Christian, A. (2017) ‘On the suppression of medical evidence’, *Journal for General Philosophy of Science*, 48(3), pp. 395–418.
- Clarke, B. *et al.* (2014) ‘Mechanisms and the Evidence hierarchy’, *Topoi*, 33(2), pp. 339–360.
- Claveau, F. (2012) ‘The Russo–Williamson theses in the social sciences: causal inference drawing on two types of evidence’, *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, 43(4), pp. 806–813.
- Cochrane, A. L. (1972) *Effectiveness and efficiency: Random reflections on health services*. London: Nuffield Provincial Hospitals Trust (The Rock Carling Fellowship, 1971).
- Cochrane Group (2021) *About us*. Available at: <https://www.cochrane.org/about-us> (Accessed: 26 April 2020).
- Cooper, R. (2002) ‘Disease’, *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, 33(2), pp. 263–282.
- Cover, T. and Hart, P. (1967) ‘Nearest neighbor pattern classification’, *IEEE Transactions on Information Theory*, 13(1), pp. 21–27.
- Craver, C. F. (2007) *Explaining the brain: Mechanisms and the mosaic unity of neuroscience*. Oxford: Oxford University Press.
- Craver, C. F. and Darden, L. (2013) *In search of mechanisms: Discoveries across the life sciences*. Chicago: The University of Chicago Press.
- Creel, K. A. (2020) ‘Transparency in complex computational systems’, *Philosophy of Science*.
- Deeks, J. J. (2002) ‘Issues in the selection of a summary statistic for meta-analysis of clinical trials with binary outcomes’, *Statistics in Medicine*, 21(11), pp. 1575–1600.
- De Fauw, J. *et al.* (2018) ‘Clinically applicable deep learning for diagnosis and referral in retinal disease’, *Nature Medicine*, 24(9), pp. 1342–1350.
- de la Monte, S. M. and Wands, J. R. (2008) ‘Alzheimer’s disease is type 3 diabetes—evidence reviewed’, *Journal of Diabetes Science and Technology*, 2(6), pp. 1101–1113.
- De Pretis, F., Landes, J. and Osimani, B. (2019) ‘E-synthesis: A Bayesian framework for causal assessment in pharmacosurveillance’, *Frontiers in Pharmacology*, 10.
- de Regt, H. W. (2017) *Understanding scientific understanding*. Oxford: Oxford University Press.

- de Regt, H. W. and Dieks, D. (2005) 'A contextual approach to scientific understanding', *Synthese*, 144(1), pp. 137–170.
- de Regt, H. W., Leonelli, S. and Eigner, K. (eds) (2009) *Scientific understanding: Philosophical perspectives*. Pittsburgh: University of Pittsburgh Press.
- Dhoro, M. *et al.* (2015) 'CYP2B6*6, CYP2B6*18, Body weight and sex are predictors of efavirenz pharmacokinetics and treatment response: population pharmacokinetic modeling in an HIV/AIDS and TB cohort in Zimbabwe', *BMC Pharmacology and Toxicology*, 16(1).
- Doherty, S. (2005) 'Evidence-based medicine: Arguments for and against', *Emergency Medicine Australasia*, 17(4), pp. 307–313.
- Doshi-Velez, F. and Kim, B. (2017) 'Towards a rigorous science of interpretable machine learning', *arXiv:1702.08608 [cs, stat]*.
- Douglas, H. (2000) 'Inductive risk and values in science', *Philosophy of Science*, 67(4), pp. 559–579.
- Dragulinescu, S. (2010) 'Diseases as natural kinds', *Theoretical Medicine and Bioethics*, 31(5), pp. 347–369.
- Dupré, J. (2020) 'Life as process', *Epistemology & Philosophy of Science*, 57(2), pp. 96–113.
- Erasmus, A., Holman, B., and Ioannidis, J.P.A. (2020) 'Data-dredging bias'. In *Catalogue of Bias*. <https://catalogofbias.org/biases/data-dredging-bias>
- Erasmus, A., Brunet, T. D. P. and Fisher, E. (2020) 'What is interpretability?', *Philosophy & Technology*.
- Esteva, A. *et al.* (2017) 'Dermatologist-level classification of skin cancer with deep neural networks', *Nature*, 542(7639), pp. 115–118.
- European Commission, Directorate-General for Research and Innovation (2019) *As predicted: Preventing p-hacking*. European Union.
- Evidence-Based Medicine Working Group (1992) 'Evidence-based medicine: A new approach to teaching the practice of medicine', *JAMA*, 268(17), p. 2420.
- Farne, H. *et al.* (2020) 'Repurposing existing drugs for the treatment of COVID-19', *Annals of the American Thoracic Society*, 17(10), pp. 1186–1194.
- Finlayson, S. G. *et al.* (2019) 'Adversarial attacks on medical machine learning', *Science*, 363(6433), pp. 1287–1289.
- Flacco, M. E. *et al.* (2015) 'Head-to-head randomized trials are mostly industry sponsored and almost always favor the industry sponsor', *Journal of Clinical Epidemiology*, 68(7), pp. 811–820.
- Fleming, N. (2018) 'Computer-calculated compounds', *Nature*, 557, pp. 555–557.
- Fleming, P. S. *et al.* (2015) 'Outcome discrepancies and selective reporting: Impacting the leading journals?', *PLOS ONE*. Edited by I. Boutron, 10(5), p. e0127495.
- Fletcher, S. C., Landes, J. and Poellinger, R. (2019) 'Evidence amalgamation in the sciences: An introduction', *Synthese*, 196(8), pp. 3163–3188.

- Food and Drug Administration (2007) 'Food and Drug Administration Amendments Act of 2007', *Public Law* 110–185. (<http://www.gpo.gov/fdsys/pkg/PLAW-110publ85/pdf/PLAW-110publ85.pdf>)
- Forstmeier, W., Wagenmakers, E.-J. and Parker, T. H. (2017) 'Detecting and avoiding likely false-positive findings - a practical guide: Avoiding false-positive findings', *Biological Reviews*, 92(4), pp. 1941–1968.
- Frosst, N. and Hinton, G. (2017) 'Distilling a neural network into a soft decision tree', *arXiv:1711.09784 [cs, stat]*.
- Fuller, J. (2018a) 'Meta-research evidence for evaluating therapies', *Philosophy of Science*, 85(5), pp. 767–780.
- Fuller, J. (2018b) 'Universal etiology, multifactorial diseases and the constitutive model of disease classification', *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, pp. 8–15.
- Fuller, J. (2021) 'The myth and fallacy of simple extrapolation in medicine', *Synthese*, 198, pp. 2919–2939.
- Fuller, J. and Flores, L. J. (2015) 'The Risk GP Model: The standard model of prediction in medicine', *Studies in History and Philosophy of Science Part C*, 54, pp. 49–61.
- Furukawa, T. A., Guyatt, G. H. and Griffith, L. E. (2002) 'Can we individualize the “number needed to treat”? An empirical study of summary effect measures in meta-analyses', *International Journal of Epidemiology*, 31(1), pp. 72–76.
- Gelman, A. (2004) 'Exploratory data analysis for complex models', *Journal of Computational and Graphical Statistics*, 13(4), pp. 755–779.
- Gelman, A. and Loken, E. (2014) 'The statistical crisis in science', *American Scientist*, 102(6), p. 460.
- Ghaemi, S. N. (2012) 'Taking disease seriously: beyond “pragmatic” nosology', in Kendler, K. S. and Parnas, J. (eds) *Philosophical Issues in Psychiatry II*. Oxford University Press, pp. 42–53.
- Ghaemi, S. N. (2013) 'Taking disease seriously in DSM', *World Psychiatry*, 12(3), pp. 210–212.
- Gillies, D. (2017) 'Mechanisms in medicine', *Axiomathes*, 27(6), pp. 621–634.
- Gillies, D. (2018) *Causality, probability, and medicine*. 1st Edition. New York: Routledge.
- Glasziou, P. P. and Irwig, L. M. (1995) 'An evidence based approach to individualising treatment', *BMJ*, 311(7016), pp. 1356–1359.
- Glennan, S. (2002) 'Rethinking mechanistic explanation', *Philosophy of Science*, 69(S3), pp. S342–S353.
- Godfrey-Smith, P. (2016) 'Mind, matter, and metabolism', *Journal of Philosophy*, 113(10), pp. 481–506.
- Goldacre, B. (2012) *Bad pharma: How drug companies mislead doctors and harm patients*. London: Fourth Estate.

- Goldacre, B., Drysdale, H., & Powell-Smith, A., et al. (2016) *The COMPare trials project*.
www.COMPare-trials.org.
- Gonthier, G. (2005) 'A computer-checked proof of the Four Colour Theorem'. Available at:
<http://research.microsoft.com/~gonthier/4colproof.pdf>, 2005.
- Gonzales, J. E. and Cunningham, C. A. (2015) *The promise of pre-registration in psychological research, American Psychological Association*. Available at:
<https://www.apa.org/science/about/psa/2015/08/pre-registration> (Accessed: 16 November 2020).
- Goodfellow, I., Bengio, Y. and Courville, A. (2016) *Deep learning*. Cambridge, Massachusetts: The MIT Press (Adaptive computation and machine learning).
- Gøtzsche, P. C. (2013) *Deadly medicines and organised crime: how big pharma has corrupted healthcare*. London: Radcliffe Publishing.
- GRADE Working Group. (2013) *GRADE handbook*. H. Schünemann, J. Brożek, G. Guyatt, and A. Oxman, eds. <https://gdt.gradepro.org/app/handbook/handbook.html#h.buaodtl66dyx>
- Greene, H. L. et al. (1992) 'The cardiac arrhythmia suppression trial: First CAST ... then CAST-II', *Journal of the American College of Cardiology*, 19(5), pp. 894–898.
- Gunning, D. (2017) 'Explainable artificial intelligence (XAI)', p. 36.
- Guyatt, G. et al. (eds) (2015) *Users' guides to the medical literature. Essentials of evidence-based clinical practice*. Third edition. New York: McGraw-Hill Education Medical.
- Guyatt, G. H. et al. (2008) 'GRADE: an emerging consensus on rating quality of evidence and strength of recommendations', *BMJ*, 336(7650), pp. 924–926.
- Hawkes, N. (2018) 'Sixty seconds on . . . p-hacking', *BMJ*, p. k4039.
- Head, M. L. et al. (2015) 'The extent and consequences of p-hacking in science', *PLOS Biology*, 13(3), p. e1002106.
- Hempel, C. G. (1965) *Aspects of scientific explanation and other essays in the philosophy of science*. New York: The Free Press.
- Hempel, C. G. and Oppenheim, P. (1948) 'Studies in the logic of explanation', *Philosophy of Science*, 15(2), pp. 135–175.
- Hitzig, Z. and Stegenga, J. (2020) 'The problem of new evidence: P-hacking and pre-analysis plans', *Diametros*, pp. 1–24.
- Hlatky, M. A. et al. (2009) 'Coronary artery bypass surgery compared with percutaneous coronary interventions for multivessel disease: a collaborative analysis of individual patient data from ten randomised trials', *Lancet*, 373(9670), pp. 1190–1197.
- Holman, B. (2019) 'In defense of meta-analysis', *Synthese*, 196(8), pp. 3189–3211.
- Hopewell, S. et al. (2009) 'Publication bias in clinical trials due to statistical significance or direction of trial results', *Cochrane Database of Systematic Reviews*. Edited by Cochrane Methodology Review Group. Available at:
<http://doi.wiley.com/10.1002/14651858.MR000006.pub3> (Accessed: 21 March 2021).

- Hopkins, S. E. (2017) 'Acute flaccid myelitis: Etiologic challenges, diagnostic and management considerations', *Current Treatment Options in Neurology*, 19(12). Available at: <http://link.springer.com/10.1007/s11940-017-0480-3> (Accessed: 10 June 2018).
- Horby, P. *et al.* (2020) *Effect of dexamethasone in hospitalized patients with COVID-19: Preliminary report*. preprint. Infectious Diseases (except HIV/AIDS).
- Hornik, K., Stinchcombe, M. and White, H. (1989) 'Multilayer feedforward networks are universal approximators', *Neural Networks*, 2(5), pp. 359–366.
- Horton, R. (2015) 'Offline: What is medicine's 5 sigma?', *Lancet*, 85(9976): 1380.
- Howard, B. *et al.* (2017) 'Systematic review: Outcome reporting bias is a problem in high impact factor neurology journals', *PLOS ONE*, 12(7), p. e0180986.
- Howick, J. (2011a) 'Exposing the vanities—and a qualified defense—of mechanistic reasoning in health care decision making', *Philosophy of Science*, 78(5), pp. 926–940.
- Howick, J. (2011b) *The philosophy of evidence-based medicine*. Chichester, West Sussex, UK: Wiley-Blackwell, BMJ Books.
- Howick, J., Glasziou, P. and Aronson, J. K. (2010) 'Evidence-based mechanistic reasoning', *Journal of the Royal Society of Medicine*, 103(11), pp. 433–441.
- Hrobjartsson, A. *et al.* (2012) 'Observer bias in randomised clinical trials with binary outcomes: systematic review of trials with both blinded and non-blinded outcome assessors', *BMJ*, 344(feb27 2), pp. e1119–e1119.
- IARC. (2019), *Preamble to the IARC monographs on the evaluation of carcinogenic risks in humans*. <https://monographs.iarc.fr/wp-content/uploads/2019/07/Preamble-2019.pdf>.
- Illari, P. M. (2011) 'Mechanistic evidence: Disambiguating the Russo–Williamson thesis', *International Studies in the Philosophy of Science*, 25(2), pp. 139–157.
- Ioannidis, J. P. A. (2005) 'Why most published research findings are false', *PLoS Medicine*, 2(8), p. e124.
- Ioannidis, J. P. A. (2008) 'Why most discovered true associations are Inflated', *Epidemiology*, 19(5), pp. 640–648.
- Ioannidis, J. P. A. (2019) 'What have we (not) learnt from millions of scientific papers with p-values?', *The American Statistician*, 73(sup1), pp. 20–25.
- Ioannidis, J. P. and Trikalinos, T. A. (2007) 'An exploratory test for an excess of significant findings', *Clinical Trials*, 4(3), pp. 245–253.
- Jacobson, H. (1959) 'The informational content of mechanisms and circuits', *Information and Control*, 2(3), pp. 285–296.
- Jager, L. R. and Leek, J. T. (2014) 'An estimate of the science-wise false discovery rate and application to the top medical literature', *Biostatistics*, 15(1), pp. 1–12.
- Jukola, S. (2015) 'Meta-Analysis, Ideals of Objectivity, and the Reliability of Medical Knowledge', *Science & Technology Studies*, 28(3), pp. 101–120.

- Jukola, S. (2017) 'On ideals of objectivity, judgments, and bias in medical research – A comment on Stegenga', *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, 62, pp. 35-41.
- Kandimalla, R., Thirumala, V. and Reddy, P. H. (2017) 'Is Alzheimer's disease a type 3 diabetes? A critical appraisal', *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease*, 1863(5), pp. 1078–1089.
- Karimi, A.H., Schölkopf, B., & Valera, I. (2020) 'Algorithmic recourse: from counterfactual explanations to interventions', *arXiv:2002.06278*.
- Kasper, D., Fauci, A., Hauser, S., Longo, D., Larry Jameson, J., & Loscalzo, J. (eds.) (2015) *Harrison's Principles of Internal Medicine*, 19th Edition. New York: McGraw-Hill Education.
- Kaplan, R. M. and Irvin, V. L. (2015) 'Likelihood of null effects of large NHLBI clinical trials has increased over time', *PLOS ONE*. Edited by S. Garattini, 10(8), p. e0132382.
- Khalifa, K. (2017) *Understanding, explanation, and scientific knowledge*. New York: Cambridge University Press.
- Kirkham, J. J., Altman, D. G. and Williamson, P. R. (2010) 'Bias due to changes in specified outcomes during the systematic review process', *PLoS ONE*. Edited by A. Vickers, 5(3), p. e9810.
- Klement, E. P., Koczy, L. T. and Moser, B. (1999) 'Are fuzzy systems universal approximators?', *International Journal of General Systems*, 28(2–3), pp. 259–282.
- Kosko, B. (1994) 'Fuzzy systems as universal approximators', *IEEE Transactions on Computers*, 43(11), pp. 1329–1333.
- Krieger, N. (1994) 'Epidemiology and the web of causation: Has anyone seen the spider?', *Social Science & Medicine*, 39(7), pp. 887–903.
- Krishnan, M. (2019) 'Against interpretability: A critical examination of the interpretability problem in machine learning', *Philosophy & Technology*.
- Kvanvig, J.L. (2009) 'The Value of Understanding', in A. Haddock, A. Millar, & D. Pritchards (eds.), *Epistemic Value*. Oxford: Oxford University Press.
- Lage, I. *et al.* (2018) 'An evaluation of the human-interpretability of explanation', p. 7.
- Landes, J., Osimani, B. and Poellinger, R. (2018) 'Epistemology of causal inference in pharmacology: Towards a framework for the assessment of harms', *European Journal for Philosophy of Science*, 8(1), pp. 3–49.
- Lehman, C. D. *et al.* (2019) 'Mammographic breast density assessment using deep learning: Clinical implementation', *Radiology*, 290(1), pp. 52–58.
- Levy, A. (2013) 'Three kinds of new mechanism', *Biology & Philosophy*, 28(1), pp. 99–114.
- Lexchin J, Bero L., Djulbegovic, B., and Clark, O. (2003) 'Pharmaceutical industry sponsorship and research outcome and quality: Systematic review', *BMJ*, 326:1167-70.

- Lindsay, R. and Cosman, F. (2015) ‘Osteoporosis’, in Kasper, D. et al. (eds) *Harrison’s Principles of Internal Medicine*. 19th Edition. New York, NY: McGraw-Hill Education.
- Lipton, P. (2009) ‘Understanding without explanation’, in de Regt, H. W., Leonelli, S., and Eigner, K. (eds) *Scientific Understanding: Philosophical Perspectives*. University of Pittsburgh Press, pp. 43–63.
- Lipton, Z. C. (2018) ‘The mythos of model interpretability’, *Queue*, 16(3), pp. 31–57.
- London, A. J. (2019) ‘Artificial intelligence and black-box medical decisions: Accuracy versus explainability’, *Hastings Center Report*, 49(1), pp. 15–21.
- Lundberg, S. and Lee, S.-I. (2017) ‘A unified approach to interpreting model predictions’, *arXiv:1705.07874 [cs, stat]*.
- Lundh, A. et al. (2012) ‘Industry sponsorship and research outcome’, *Cochrane Database of Systematic Reviews*.
- Lundh, A. et al. (2017) ‘Industry sponsorship and research outcome’, *Cochrane Database of Systematic Reviews*.
- Machamer, P., Darden, L. and Craver, C. F. (2000) ‘Thinking about mechanisms’, *Philosophy of Science*, 67(1), pp. 1–25.
- Mackie, J. L. (1965) ‘Causes and conditions’, *American Philosophical Quarterly*, 2(4), pp. 245–264.
- MacMahon, B., Pugh, T. F. and Ipsen, J. (1960) *Epidemiologic methods*. Boston: Little and Brown Company.
- Madadi, P. et al. (2013) ‘Clinical practice guideline: CYP2D6 genotyping for safe and efficacious codeine therapy’, *Journal of Population Therapeutics and Clinical Pharmacology*, 20(3), pp. e369-396.
- Maimbo, M. et al. (2012) ‘CYP2B6 genotype is a strong predictor of systemic exposure to efavirenz in HIV-infected Zimbabweans’, *European Journal of Clinical Pharmacology*, 68(3), pp. 267–271.
- Marchionni, C. and Reijula, S. (2019) ‘What is mechanistic evidence, and why do we need it for evidence-based policy?’, *Studies in History and Philosophy of Science Part A*, 73, pp. 54–63.
- McDonald, V.M. et al. (2019) ‘Treatable traits can be identified in a severe asthma registry and predict future exacerbations’, *Respirology*, 24(1), pp. 37–47.
- McKinney, S. M. et al. (2020) ‘International evaluation of an AI system for breast cancer screening’, *Nature*, 577(7788), pp. 89–94.
- Miller, T. (2019) ‘Explanation in artificial intelligence: Insights from the social sciences’, *Artificial Intelligence*, 267, pp. 1–38.
- Mishra, S., Sturm, B. L. and Dixon, S. (2017) ‘Local interpretable model-agnostic explanations for music content analysis’, *ISMIR*, pp. 537–543.

- Mittelstadt, B. D. *et al.* (2016) 'The ethics of algorithms: Mapping the debate', *Big Data & Society*, 3(2).
- Mittelstadt, B., Russell, C. and Wachter, S. (2019) 'Explaining explanations in AI', in *Proceedings of the Conference on Fairness, Accountability, and Transparency - FAT* '19. the Conference*, Atlanta, GA, USA: ACM Press, pp. 279–288.
- Moher, D. *et al.* (2010) 'CONSORT 2010 explanation and elaboration: Updated guidelines for reporting parallel group randomised trials', *BMJ*, 340, pp. c869–c869.
- Munafò, M. R. *et al.* (2017) 'A manifesto for reproducible science', *Nature Human Behaviour*, 1, p. 0021.
- Murad, M. H. *et al.* (2018) 'The effect of publication bias magnitude and direction on the certainty in evidence', *BMJ Evidence-Based Medicine*, 23(3), pp. 84–86.
- Narayanan, M. *et al.* (2018) 'How do humans understand explanations from machine learning systems? An evaluation of the human-interpretability of explanation', *arXiv:1802.00682 [cs, stat]*.
- National Institutes of Health, Department of Health and Human Services. (2016) 'Clinical trials registration and results information submission: Final rule', *Fed Regist 2016*, 81:64981-65157.
- National Institute for Health Research (2019) *Mechanistic studies, explanation and examples*. Available at: <https://www.nihr.ac.uk/documents/mechanistic-studies-explanation-and-examples/12146?pr=> (Accessed: 29 January 2020).
- New Scientist (2020) *Covid-19 shows why an infodemic of bad science must never happen again*, *New Scientist*. Available at: <https://institutions.newscientist.com/article/mg24632812-500-covid-19-shows-why-an-infodemic-of-bad-science-must-never-happen-again/> (Accessed: 16 November 2020).
- Nordling, L. (2017) 'Putting genomes to work in africa', *Nature*, 544, pp. 20–22.
- Nosek, B. A., Spies, J. R., and Motyl, M. (2012) 'Scientific utopia, II: Restructuring incentives and practices to promote truth over publishability', *Perspectives on Psychological Science*, 7, pp. 615–31.
- Nuijten, M. B. *et al.* (2016) 'The prevalence of statistical reporting errors in psychology (1985–2013)', *Behavior Research Methods*, 48(4), pp. 1205–1226.
- Nutescu, E. A. *et al.* (2016) 'Pharmacology of anticoagulants used in the treatment of venous thromboembolism', *Journal of Thrombosis and Thrombolysis*, 41(1), pp. 15–31.
- Nyakutira, C. *et al.* (2008) 'High prevalence of the CYP2B6 516G→T(*6) variant and effect on the population pharmacokinetics of efavirenz in HIV/AIDS outpatients in Zimbabwe', *European Journal of Clinical Pharmacology*, 64(4), pp. 357–365.
- Obermeyer, Z. *et al.* (2019) 'Dissecting racial bias in an algorithm used to manage the health of populations', *Science*, 366(6464), pp. 447–453.

- OCEBM Levels of Evidence Working Group (2011) 'The Oxford 2011 levels of evidence', *Oxford Centre for Evidence-Based Medicine*. Available at: <https://www.cebm.net/index.aspx?o=5653>.
- Odgaard-Jensen, J. et al. (2011) 'Randomisation to protect against selection bias in healthcare trials', *Cochrane Database of Systematic Reviews*. Edited by Cochrane Methodology Review Group.
- O'Gara, P. T. and Loscalzo, J. (2015) 'Mitral valve disease', in Kasper, D. et al. (eds) *Harrison's Principles of Internal Medicine*. 19th Edition. New York, NY: McGraw-Hill Education.
- Open Science Collaboration (2015) 'Estimating the reproducibility of psychological science', *Science*, 349(6251), p. aac4716.
- Páez, A. (2019) 'The pragmatic turn in explainable artificial intelligence (XAI)', *Minds and Machines*, 29(3), pp. 441–459.
- Park, A., D. Steel, and E. Maine. (*forthcoming*) 'Evidence-based medicine and mechanistic evidence: The case of the failed rollout of efavirenz in Zimbabwe', *Journal of Medicine and Philosophy*.
- Parkkinen, V.-P. (2016) 'Robustness and evidence of mechanisms in early experimental atherosclerosis research', *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, 60, pp. 44–55.
- Parkkinen, V.-P. et al. (2018) *Evaluating evidence of mechanisms in medicine: Principles and procedures*. Cham: Springer International Publishing.
- Peacock, S. J. and Paterson, G. K. (2015) 'Mechanisms of methicillin resistance in *Staphylococcus aureus*', *Annual Review of Biochemistry*, 84(1), pp. 577–601.
- Pearl, J., and Bareinboim, E. (2014) 'External Validity: From Do-Calculus to Transportability Across Populations', *Statistical Science* 29(4), pp. 579–595.
- Perneger, T. V. and Combescure, C. (2017) 'The distribution of p-values in medical research articles suggested selective reporting associated with statistical significance', *Journal of Clinical Epidemiology*, 87, pp. 70–77.
- Pignon, J.-P. et al. (2009) 'Meta-analysis of chemotherapy in head and neck cancer (MACH-NC): an update on 93 randomised trials and 17,346 patients', *Radiotherapy and Oncology: Journal of the European Society for Therapeutic Radiology and Oncology*, 92(1), pp. 4–14.
- Post, P. N., de Beer, H. and Guyatt, G. H. (2013) 'How to generalize efficacy results of randomized trials: recommendations based on a systematic review of possible approaches: Generalizing efficacy results of RCTs', *Journal of Evaluation in Clinical Practice*, 19(4), pp. 638–643.
- Potochnik, A. (2016) 'Scientific explanation: Putting communication first', *Philosophy of Science*, 83(5), pp. 721–732.

- Prior, M. *et al.* (2017) 'Inadvertent p-hacking among trials and systematic reviews of the effect of progestogens in pregnancy? A systematic review and meta-analysis', *BJOG: An International Journal of Obstetrics & Gynaecology*, 124(7), pp. 1008–1015.
- Rajkumar, A. *et al.* (2018) 'Ensuring fairness in machine learning to advance health equity', *Annals of Internal Medicine*, 169(12), p. 866.
- Rajpurkar, P. *et al.* (2017) 'CheXNet: Radiologist-level pneumonia detection on chest x-rays with deep learning', *arXiv:1711.05225 [cs, stat]*.
- Raviglione, M. C. (2015) 'Tuberculosis', in Kasper, D. *et al.* (eds) *Harrison's Principles of Internal Medicine*. 19th Edition. New York, NY: McGraw-Hill Education.
- Rehfuess, E. A. *et al.* (2016) 'An approach for setting evidence-based and stakeholder-informed research priorities in low- and middle-income countries', *Bulletin of the World Health Organization*, 94(4), pp. 297–305.
- Rezende, L. F. M. de *et al.* (2018) 'Reporting bias in the literature on the associations of health-related behaviors and statins with cardiovascular disease and all-cause mortality', *PLOS Biology*. Edited by M. Macleod, 16(6), e2005761.
- Ribeiro, M. T., Singh, S. and Guestrin, C. (2016) "“Why should I trust you?”: Explaining the predictions of any classifier', *arXiv:1602.04938 [cs, stat]*.
- Ribeiro, M. T., Singh, S. and Guestrin, C. (2018) 'Anchors: High-precision model-agnostic explanations', *AAAI Conference on Artificial Intelligence; Thirty-Second AAAI Conference on Artificial Intelligence*.
- Ridker, P. M. and Torres, J. (2006) 'Reported Outcomes in Major Cardiovascular Clinical Trials Funded by For-Profit and Not-for-Profit Organizations: 2000-2005', *JAMA*, 295(19), p. 2270.
- Rothman, K. J. (1976) 'CAUSES', *American Journal of Epidemiology*, 104(6), pp. 587–592.
- Rudner, R. (1953) 'The scientist qua scientist makes value judgments', *Philosophy of Science*, 20(1), pp. 1–6.
- Russo, F. and Williamson, J. (2007) 'Interpreting causality in the health sciences', *International Studies in the Philosophy of Science*, 21(2), pp. 157–170.
- Salmon, W. (1971) 'Statistical explanation', in Salmon, W. (ed.) *Statistical Explanation & Statistical Relevance*. Pittsburgh: University of Pittsburgh Press, pp. 29–87.
- Salmon, W. (1984) *Scientific explanation and the causal structure of the world*. Princeton: Princeton University Press.
- Salmon, W. (1989) *Four decades of scientific explanation*. Pittsburgh: University of Pittsburgh Press.
- Scadding, J. G. (1959) 'Principles of definition in medicine with special reference to chronic bronchitis and emphysema', *Lancet*, 1(7068), pp. 323–325.
- Scheirer, W. (2020) 'A pandemic of bad science', *Bulletin of the Atomic Scientists*, 76(4), pp. 175–184.

- Schmid, C. H. et al. (1998) 'An empirical study of the effect of the control rate as a predictor of treatment efficacy in meta-analysis of clinical trials', *Statistics in Medicine*, 17(17), pp. 1923–1942.
- Sennvik, K. et al. (2000) 'Levels of α - and β -secretase cleaved amyloid precursor protein in the cerebrospinal fluid of Alzheimer's disease patients', *Neuroscience Letters*, 278(3), pp. 169–172.
- Shah, K. et al. (2020) 'Outcome reporting bias in Cochrane systematic reviews: a cross-sectional analysis', *BMJ Open*, 10(3), p. e032497.
- Simmons, J. P., Nelson, L. D. and Simonsohn, U. (2011) 'False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant', *Psychological Science*, 22(11), pp. 1359–1366.
- Simmons, J. P., Nelson, L. D. and Simonsohn, U. (2017) 'How to properly preregister a study', *Data Colada*. Available at: <http://datacolada.org/64> (Accessed: 16 November 2020).
- Simonsohn, U., Nelson, L. D. and Simmons, J. P. (2014) 'P-curve: A key to the file-drawer.', *Journal of Experimental Psychology: General*, 143(2), pp. 534–547.
- Simonsohn, U., Simmons, J. P. and Nelson, L. D. (2015) 'Better P-curves: Making P-curve analysis more robust to errors, fraud, and ambitious P-hacking, a Reply to Ulrich and Miller (2015)', *Journal of Experimental Psychology: General*, 144(6), pp. 1146–1152.
- Sismondo, S. (2008) 'How pharmaceutical industry funding affects trial outcomes: Causal structures and responses', *Social Science & Medicine*, 66(9), pp. 1909–1914.
- Smart, B. (2014) 'On the classification of diseases', *Theoretical Medicine and Bioethics*, 35(4), pp. 251–269.
- Smolik, P. (1999) 'Validity of nosological classification', *Dialogues in Clinical Neuroscience*, 1(3), pp. 185–190.
- Snider, G. L. (2003) 'Nosology for our day: its application to chronic obstructive pulmonary disease', *American Journal of Respiratory and Critical Care Medicine*, 167(5), pp. 678–683.
- Sober, E. (2007) 'Evidence and Value Freedom', in H. Kincaid, J. Dupré, and A. Wylie (eds.) *Value-Free Science?* Oxford: Oxford University Press.
- Solomon, M. (2011) 'Just a paradigm: evidence-based medicine in epistemological context', *European Journal for Philosophy of Science*, 1(3), pp. 451–466.
- Solomon, M. (2015) *Making medical knowledge*. First edition. Oxford: Oxford University Press.
- Somashekhar, S. P. et al. (2018) 'Watson for oncology and breast cancer treatment recommendations: Agreement with an expert multidisciplinary tumor board', *Annals of Oncology*, 29(2), pp. 418–423.
- Steel, D. (2008) *Across the boundaries: Extrapolation in biology and social science*. Oxford: Oxford University Press (Environmental ethics and science policy series).

- Steel, D. (2018) 'If the facts were not untruths, their implications were: Sponsorship bias and misleading communication', *Kennedy Institute of Ethics Journal*, 28(2), pp. 119–144.
- Stegenga, J. (2011) 'Is meta-analysis the platinum standard of evidence?', *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, 42(4), pp. 497–507.
- Stegenga, J. (2013) 'An impossibility theorem for amalgamating evidence', *Synthese*, 190(12), pp. 2391–2411.
- Stegenga, J. (2014) 'Down with the hierarchies', *Topoi*, 33(2), pp. 313–322.
- Stegenga, J. (2015a) 'Effectiveness of medical interventions', *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, 54, pp. 34–44.
- Stegenga, J. (2015b) 'Herding QATs: Quality assessment tools for evidence in medicine', in Huneman, P., Lambert, G., and Silberstein, M. (eds) *Classification, Disease and Evidence*. Dordrecht: Springer Netherlands, pp. 193–211.
- Stegenga, J. (2015c) 'Measuring effectiveness', *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, 54, pp. 62–71.
- Stegenga, J. (2018) *Medical nihilism*. New York: Oxford University Press.
- Stegenga, J. (n.d.) 'Bayesian Mechanista', *Unpublished manuscript*.
- Stegenga, J. and Menon, T. (2017) 'Robustness and independent evidence', *Philosophy of Science*, 84(3), pp. 414–435.
- Stone, C. J. (1977) 'Consistent nonparametric regression', *The Annals of Statistics*, 5(4), pp. 595–620.
- Straus, S. E. *et al.* (eds) (2018) *Evidence-based medicine: how to practice and teach EBM*. London: Elsevier.
- Strevens, M. (2011) *Depth: an account of scientific explanation*. 1st Edition. Cambridge, MA: Harvard Univ. Press.
- Tabatabaei Ghomi, H., and Stegenga, J. (n.d.) 'Computational investigation of methodological bias in antidepressant RCTs', *Unpublished manuscript*.
- The RECOVERY Collaborative Group (2020) 'Dexamethasone in hospitalized patients with COVID-19 — preliminary report', *New England Journal of Medicine*.
- Toulmin, S. E. (2003) *The uses of argument*. 2nd Edition. Cambridge: Cambridge University Press.
- Tschandl, P. *et al.* (2019) 'Comparison of the accuracy of human readers versus machine-learning algorithms for pigmented skin lesion classification: an open, web-based, international, diagnostic study', *The Lancet Oncology*, 20(7), pp. 938–947.
- Turnbull, Fiona *et al.* (2008) 'Do men and women respond differently to blood pressure-lowering treatment? Results of prospectively designed overviews of randomized trials', *European Heart Journal*, 29(21), pp. 2669–2680.

- Turnbull, F. et al. (2008) 'Effects of different regimens to lower blood pressure on major cardiovascular events in older and younger adults: meta-analysis of randomised trials', *BMJ* (Clinical research ed.), 336(7653), pp. 1121–1123.
- Upshur, R. E. G. (2005) 'Looking for rules in a world of exceptions: Reflections on evidence-based practice', *Perspectives in Biology and Medicine*, 48(4), pp. 477–489.
- Valles, S. A. (2018) *Philosophy of population health science: philosophy for a new public health era*. London: Routledge.
- van Fraassen, Bas. C. (1980) *The scientific image*. Oxford University Press.
- Vera-Badillo, F. E. et al. (2016) 'Bias in reporting of randomised clinical trials in oncology', *European Journal of Cancer*, 61, pp. 29–35.
- Waldner, D. (2012) 'Process tracing and causal mechanisms', in Kincaid, H. (ed.) *The Oxford Handbook of Philosophy of Social Science*. Oxford: Oxford University Press, pp. 65–84.
- Walker, B. R. et al. (eds) (2014) *Davidson's principles and practice of medicine*. 22nd edition. New York: Elsevier.
- Wang, L.-X. (1992) 'Fuzzy systems are universal approximators', in [1992 Proceedings] *IEEE International Conference on Fuzzy Systems*. [1992 Proceedings] *IEEE International Conference on Fuzzy Systems*, San Diego, CA, USA: IEEE, pp. 1163–1170.
- Wasserstein, R. L. and Lazar, N. A. (2016) 'The ASA statement on p-values: Context, process, and purpose', *The American Statistician*, 70(2), pp. 129–133.
- Watson, D. S. et al. (2019) 'Clinical applications of machine learning algorithms: beyond the black box', *BMJ*, p. 1886.
- Weisberg, M. (2006) 'Robustness analysis', *Philosophy of Science*, 73(5), pp. 730–742.
- Whitbeck, C. (1977) 'Causation in Medicine: The Disease Entity Model', *Philosophy of Science*, 44(4), pp. 619–637.
- Whitehead, A. N., (1929) [1985]. *Process and Reality*, New York: Macmillan. Corrected edition, D.R. Griffin & D. W. Sherburne (eds.), New York: The Free Press, 1985.
- Wiens, J. and Shenoy, E. S. (2018) 'Machine learning for healthcare: On the verge of a major shift in healthcare epidemiology', *Clinical Infectious Diseases*, 66(1), pp. 149–153.
- Woodward, J. (1989) 'The causal/mechanical model of explanation', in Kitcher, P. and Salmon, W. (eds) *Scientific Explanation*. (Minnesota Studies in the Philosophy of Science, 13), pp. 357–383.
- Woodward, J. (2003) *Making things happen*. Oxford University Press.
- Woodward, J. (2006) 'Some Varieties of Robustness', *Journal of Economic Methodology*, 13, pp. 219–40.
- Woodward, J. (2017) 'Scientific explanation', in Zalta, E. N. (ed.) *The Stanford Encyclopedia of Philosophy*. Fall 2017. Metaphysics Research Lab, Stanford University. Available at: <https://plato.stanford.edu/archives/fall2017/entries/scientific-explanation>.

- World Health Organization (2016) *International statistical classification of diseases and related health problems*.
- World Health Organization (n.d.) *WHO / International Classification of Diseases, WHO*. Available at: <http://www.who.int/classifications/icd/en/> (Accessed: 2 June 2018).
- Worrall, J. (2002) ‘What evidence in evidence-based medicine?’, *Proceedings of the Philosophy of Science Association*, 2002(3), pp. S316–S330.
- Wüthrich, N. and Steele, K. (2019) ‘The problem of evaluating automated large-scale evidence aggregators’, *Synthese*, 196(8), pp. 3083–3102.
- Yelle, D. (2012) *Ischemic heart disease / McMaster pathophysiology review*. Available at: <http://www.pathophys.org/acs/> (Accessed: 4 May 2018).
- Yen, J., Liang Wang and Gillespie, C. W. (1998) ‘Improving the interpretability of TSK fuzzy models by combining global learning and local learning’, *IEEE Transactions on Fuzzy Systems*, 6(4), pp. 530–537.
- Zachar, P. and Kendler, K. S. (2017) ‘The philosophy of nosology’, *Annual Review of Clinical Psychology*, 13(1), pp. 49–71.
- Zadeh, L. A. (1965) ‘Fuzzy sets’, *Information and Control*, 8(3), pp. 338–353.
- Zednik, C. (2019) ‘Solving the black box problem: A normative framework for explainable artificial intelligence’, *Philosophy & Technology*.
- Zerilli, J. *et al.* (2019) ‘Transparency in algorithmic and human decision-making: Is there a double standard?’, *Philosophy & Technology*, 32(4), pp. 661–683.
- Zhang, T. *et al.* (2019) ‘ANODEV2: A coupled neural ODE evolution framework’, *arXiv:1906.04596 [cs, stat]*.
- Zhou, Y., Zhou, Z. and Hooker, G. (2018) ‘Approximation trees: Statistical stability in model distillation’, *arXiv:1808.07573 [cs, stat]*.