The power of real-world evidence: A critical study of evolving evidence standards for medical knowledge

Egli, Michaela

Thèse de doctorat
Université de Genève
Département de Philosophie

The power of real-world evidence

A critical study of evolving evidence standards for medical knowledge

Michaela Egli

February 2024

Supervisors
Prof. Marcel Weber
Prof. Jacob Stegenga

UNIVERSITÉ DE GENÈVE

FACULTÉ DES LETTRES

IMPRIMATUR

DOCTORAT ÈS LETTRES

Philosophie

Thèse de Michaela EGLI

Intitulée : « The power of real-word evidence: A critical study of evolving evidence standards for medical knowledge »

*

La Faculté des lettres, sur le préavis d'une commission composée de Madame et Messieurs les Professeur-es, Christian Wüthrich, président du jury ; Marcel Weber, directeur de thèse ; Jacob Stegenga, co-directeur de thèse (University of Cambridge) ; Lara Keuck (Universität Bielefeld) ; David Teira (Universidad Nacional de Educación a Distancia, Madrid), autorise l'impression de la présente thèse, sans exprimer d'opinion sur les propositions qui y sont énoncées.

Genève, le 24 juin 2024

La Doyenne : Francesca SERRA

UNIVERSITÉ de GENÈVE FACULTÉ des LETTRES

Thèse N° 1139

N.B. – La thèse doit porter la déclaration précédente * et remplir les conditions énumérées dans les informations pour la publication d'une thèse.

Un exemplaire numérisé doit être remis à la Direction de l'information scientifique.

Table of contents

List of tables

List of figures

Acknowledgements

This thesis has benefited greatly from the inspiring insights, constructive criticism and generous support of others. I would like to thank them wholeheartedly. I am grateful to David Teira for helpful and constructive discussions on several chapters of this thesis. I would like to thank all the participants of the lgBIG who read and critically discussed drafts of many chapters of this thesis and supported me in many ways over the last four years. I am also grateful to the 'Stegengasaurs' for their enlightening discussions on drafts of several chapters and for welcoming me into their group. I am grateful to Servan Grueninger for his passion for statistics and for convincing me that philosophers have got randomisation all wrong. I am grateful to Lorenzo Hess of Swissmedic for stimulating discussions, his insistence that 'it is all about GCP' and the opportunity to present my case study on the FDA approval of Prograf at the Swissmedic lunch talks. I would like to thank Lars Hemkens for his inspiring enthusiasm and his insightful lessons on pragmatic trials and Sandro Christensen, who enriched my understanding of the case studies with his medical expertise. I would also like to thank my colleagues at the Swiss Clinical Trial Organisation for their support and insights over the last four years. They have taught me many lessons about clinical trials and the complexity of all the nitty-gritty details for their execution. I am also deeply grateful to my supervisors. I thank Jacob Stegenga for welcoming me to Cambridge and for his helpful feedback on the draft material of this thesis. Most of all, I would like to thank Marcel Weber for the many doors he has opened along the way, for his constructive feedback on this thesis throughout the entire process, and for always keeping the spirits high. This research was generously funded by the Swiss National Science Foundation.

Introduction

In the last decades, digital health data has sparked new hopes for better medical research. A highly controversial topic is the re-use of such data for regulating medicines and guiding treatment decisions in healthcare. This thesis studies how *real-world evidence* changes medical knowledge about the *effectiveness of medical treatments* by looking at the epistemic shifts, opportunities, and risks involved in its use. Real-world evidence is an umbrella term for evidence that has been generated from analysing non-experimental health data, such as electronic health records (EHR), pharmacy records or insurance claims data. It lends itself to various analytical methods, including observational studies, pragmatic clinical trials, or historically controlled trials. Real-world evidence is an intriguing idea that has the power to challenge the well-established gold standard – the randomised controlled trial (RCT).

Sabina Leonelli, who pioneered the philosophy of data-centric science, describes such data as embarking on a *journey*. Leonelli drew philosophers' attention to the coming avalanche of *travelling* data and the frictions and costs that accompany it. She introduced this metaphor to illustrate that 'there is nothing smooth about data journeys' and that journeys of data require planning, involve different material infrastructure and are generally fragmented and complex just like human journeys' (Leonelli 2016). This insight is not only relevant to philosophers but also to policy makers and policy advisors, including the Swiss Science and Innovation Council (SSIC). In her exploratory study on behalf of the SSIC, Leonelli pointed to various challenges involved in data journeys including concerns for technical and semantic standardisation of data for various sources, questions of ownership and value of data, or the question of sustainable funding of research infrastructures (Leonelli 2017a). After eight years and more than CHF 100 million invested in a national health data infrastructure in Switzerland, many of these challenges remain.

Data from real-world sources have now arrived on the desks of those who regulate the use of medicines. We can currently witness the uptake of such data as an accepted form of evidence into the practices, laws and norms that govern human research and the regulation of medicines. In 2016, the 21st Century Cures Act introduced the possibility for using real-world evidence for the approval of medicines and mandated the Food and Drug Administration (FDA) to evaluate

this idea (114th Congress). From the very beginning, opinions on the amendment were divided. Its advocates argued that it would accelerate and multiply therapeutic opportunities for patients, whereas its opponents warned that the easing of evidential standards might expose patients to unnecessary risks. In 2018, the FDA published its real-world evidence program, outlining a working plan on how they intend to execute the mandate (US Food and Drug Administration 2018). Since then, the agency has produced a series of guidelines on the subject and in July 2021, the FDA took a landmark step and announced the first approval of a drug exclusively based on real-world evidence (US Food and Drug Administration 2021a). The EMA already conducted a pilot program for adaptive licensing in which real-world evidence plays an essential role in 2014 (European Medicines Agency 2014). In 2021 the agency announced that they envision to enable the use of real-world evidence and establish its value for regulatory decision-making in Europe by 2025 (Arlett et al. 2021). The Swiss regulatory agency, Swissmedic, followed suit and published a brief position paper in July 2022 on their openness to consider real-world evidence in certain circumstances (Swissmedic 2022). Most recently, a globally recognised reference organisation for evidence standards in clinical research, the International Council on Harmonisation (ICH), has revised two of its core guidelines to include innovations from the real-world evidence movement (International Council for Harmonisation 2021b, 2023). The evolution seems irreversible.

Real-world evidence raises high hopes. The phrase 'real-world evidence' alludes to a widely recognised shortcoming of the RCT gold standard. As RCTs are generally conducted under highly, artificially controlled conditions, critics claim that the evidence they generate has little to say about the conditions and patients outside experiments. Contrary to this, data collected during actual health care practices promise to be more informative about 'real' treatment settings and 'real' patients. In light of the limitations of the randomised method, philosophers and clinical researchers have welcomed the promise of the new evidence regime to provide applicable and widely generalisable evidence to better inform healthcare decisions (Hemkens et al. 2016; Kalkman et al. 2017a; Borgerson 2013; Zwarenstein and Treweek 2009; Mc Cord et al. 2018). However, the phrase 'real-world evidence' is first and foremost a clever rhetorical marketing manoeuvre. It rephrases the data's greatest weakness – the absence of experimental control – as their main advantage. The rhetorical manoeuvre certainly contributed to the current success of this development.

Another driving motivation of this development is that such data is a less expensive way to produce evidence, which responds to an innovation crisis in drug development (Eichler et al. 2013; Rawlins 2004). In the view of these authors, less costly standards for regulatory decision making will lead to more knowledge (and more treatments). Wilholt has shown that concerns for resources are philosophically and epistemically relevant (Wilholt 2013, 2016). It is epistemically desirable to rely on methods that can produce more results with the same resources, so Wilholt argues, and calls this desiderata a method's 'power'. Data that travel between multiple epistemic and non-epistemic users are the epitome of a powerful research approach in Wilholt's terms. The desire for the epistemic power that such data bring has raised high hopes for more efficient drug development and exhaustiv medical knowledge about the effectiveness of treatments in health care (Mc Cord and Hemkens 2019).

Digital health data are at the centre of even larger visionary concepts, which try to highlight the potential to revolutionise healthcare and to drive progress in medical knowledge and patient care. A particularly prominent vision, the 'health data ecosystem', portrays investments in health data generation as investments in a sustainable resource for various uses within science, policy, healthcare and technology development with the potential to circulate within an interdependent network of data users and producers. Another vision with a large scope is the 'learning healthcare system', which takes the concept of a data ecosystem a step further by advocating for the continuous recording of healthcare experiences in data, which are then to be transformed into scientific knowledge that is not confined to scientific journals (or drawers) but supposed to directly inform and enrich the everyday practices of healthcare providers (Eichler et al. 2019). As such, this vision might be able to bridge the gap between clinical expertise encoded in data and scientifically accepted ways of knowing.

To make these visions real, policy makers have initiated a wide range of initiatives and large-scale investments into digital health data. These range from small local or national initiatives, such as the Swiss Personalised Health Network, Findata or the All of Us project, to large international collaborations like the Beyond 1 Million Genomes (B1MG) project, EU Darwin, or the European Health Data Space. In these developments, healthcare professionals have become an indispensable part of the data collection pipeline. Many governments are now investing in the streamlining of data collection processes *at the source*, that is, in the clinics. One such investment programme is the US Medicare and Medicaid Electronic Health Record Incentive Program (also known as 'Meaningful Use') that

rewards health care professionals for the adoption of EHR technology and (among others) for the structured recording of core medical data. Finally, patients and the public are increasingly expected to contribute to these visions and share their health data as an act of solidarity for the greater social good. The hopes are high but critical engagement with these hopes is widely absent in the debate.

The controversy of epistemic pluralism in clinical research

The uptake of travelling data into the norms of clinical research induces a shift towards evidential pluralism on two ends: it broadens the scope of the methods *and* the types of data that are deemed acceptable as potential evidence for the effectiveness of medicines. The advent of real-world data challenges the established idea that quality of evidence is largely determined hierarchically and ensured by a fixed set of rules (called Good Clinical Practice). Instead, it introduces a contextualised approach to evaluate the quality of evidence that understands the quality of evidence relative to how well a certain piece of evidence serves in achieving a particular purpose – referred to as 'fitness-for-purpose'.

The scope of this thesis is to analyse the controversial *use* of real-world evidence for assessing the *effectiveness* of medical therapies for practical decision-making contexts. This aspect of the discussion is particularly controversial as it culminates in the fundamental question whether we can and should substitute conventional randomised trials with real-world evidence for the regulation of medicines or decision-making in healthcare (Franklin and Schneeweiss 2017). To solve this controversy, stakeholders in the field have created various real-world evidence expert groups like the Get-Real consortium and launched research projects that are dedicated to the replication of such evidence (Franklin et al. 2020a; Bartlett et al. 2019; García-Albéniz et al. 2017; Wallach et al. 2021). Authorities and data infrastructures have published numerous position papers and guidelines (US Food and Drug Administration 2023; European Medicines Agency 2024). Scientists have weighed in with countless opinion pieces in scientific journals and the skills and expertise required to generate real-world evidence are taught in various webinars. Many of these resources emerged during the last four years of writing this thesis.

For decades, randomised controlled trials (RCTs) have been an indispensable method for making decisions on drug approvals and informing healthcare decisions. A large part of the success of RCTs can be traced to the US Food and Drug Administration (FDA) reliance on them since 1970 and to the advent of evidence-based medicine (EBM) in the 1990s, with its hierarchical theory

of evidence that orders methods according to their quality and puts randomised controlled trials (RCTs) near the top of the evidence hierarchy. The EBM approach was so successful that RCTs nearly replaced all other types of evidence for decision making, such as consensus conferences, mechanistic knowledge, or clinical expertise. Many philosophers have convincingly criticised this categorical use of RCTs and proposed alternative evidence standards. For example, Osimani has argued that randomised trials are inadequate for evaluating harms (Osimani 2013) and in subsequent work with Landes and Poellinger they developed an approach for evidence amalgamation that takes all forms of evidence into account (Landes et al. 2018). Cartwright has advocated for a plurality of evidence to increase the applicability of evidence outside the research context (Cartwright 2012, 2007, 2009). The movement that calls itself EBM+ defended the claim that evidence standards should consider mechanistic knowledge (the '+') in addition to statistical evidence from randomised trials (Clarke et al. 2013, 2014). Borgerson argued that we should move towards evidence that has more social value than the current gold standard (Borgerson 2013) and Stegenga criticised the hierarchies and instead advocated for the use of quality assessment tools that turn towards method tokens rather than types (Stegenga 2018, chapter 5).

I concur insofar with the critics of EBM, as I think that deliberating on evidence standards should not be limited to assessing the internal validity of the scientific method. As my study shows, the uptake of real-world evidence into the norms of clinical research goes far beyond considerations of validity. These new methods also create *conceptual shifts* in our causal knowledge, impact the *practical relevance* of evidence or *allocate resources differently*. Hence, my study illustrates that the narrow view of the quality of evidence promoted by the EBM standard cannot adequately explain what awaits us with the real-world evidence transformation. The advantage of a contextualised approach is that we can recognise the multidimensionality of 'quality' of evidence, highlight those dimensions that are most essential in a particular context, account for available resources and resource constraints and carefully tailor the choice of methods to these circumstances. However, such a contextual approach is vulnerable to misuse and comes with serious epistemic risks that potentially outweigh its epistemic benefits.

During exchanges with practitioners and stakeholders involved in this development, I learned that many stakeholders in the field are not so much concerned with the plurality of methods as they are with the plurality of data. One reason why randomised studies are so valuable as a source of evidence for the

community is that they are embedded in a large body of globally accepted rules, with decades of accumulated experience of how data should be translated into decisions. Among these are the pivotal rules of 'Good Clinical Practice' (GCP) (International Council for Harmonisation 2016). This sixty-pages set of conventions contains globally accepted rules and principles that create accountability for the integrity of data and the safety of patients in a clinical trial. At the heart of these rules is the idea that data are inextricably linked to their purpose and they closely govern how data in clinical trials are collected, stored, reported, and verified. The GCP rules are based on the logic that there is a 'sponsor' who has overall responsibility for data integrity and that patients give informed consent to participate in a trial with a particular purpose. GCP rules also introduce the idea of data monitoring and auditing as control mechanisms for the quality of data. Data that are governed by these rules are seen as a reliable and trustworthy source of evidence by the community.

What is peculiar about the use of real-world data in contrast to clinical trial data is that the former have not been made for a specific research purpose, they have – in theory – not even been made for research. The recording of such data is not constrained by the rules of GCP but driven by the diverse needs, incentives, and constraints of local healthcare contexts. Consequently, the quality of real-world data emerges as a major concern in this development. To deal with this problem, regulators began developing data quality frameworks defining criteria and suggesting strategies to assess the 'fitness-of-purpose' of data (European Medicines Agency 2022; US Food and Drug Administration 2021c, 2021b). Many data sharing networks have published their own guidance papers on these questions (Bernal-Delgado et al. 2022; NIH Pragmatic Trials Collaboratory 2014, 2024). However, bridging the gap between the highly abstract idea of fitness-for-purpose and concrete measures of data quality has proven a notoriously complex tasks with a wide lack of convergence on how to best approach it (Illari 2014). Despite the prevalence of these discussions in the methodological literature, philosophers of medicine have largely neglected the role that data handling practices play for the reliability and trustworthiness of medical knowledge. With this thesis I contribute to closing this gap.

The scope of this thesis

My philosophical study of real-world evidence and the evolution of evidence standards for medical knowledge explores the concerns, the hopes, and the trade-offs involved in using travelling data for the clinical evaluation of medical treatments. This involves careful engagement with long standing discussions in the

epistemology of medical research such as the significance of randomisation, the problem of extrapolation, conceptual and methodological issues of data quality practices, and trust in clinical research. Each chapter not only attempts to apply insights from existing philosophical literature to understand 'real-world evidence' as a special case but also aims to advance these debates by bringing back the lessons we can learn from this development. My in-depth discussion of the increasingly popular pragmatic trial method and the epistemic importance of Good Clinical Practice (GCP) also covers genuinely new ground. I closely analyse pragmatic clinical trials, their aptness to generate valid causal conclusions and the foundations of their perceived social value. My inquiry into the epistemic value of GCP highlights the importance of these rules for both the reliability and the trustworthiness of data that philosophers of medicine hitherto largely neglected.

The thesis is structured into two parts. In Part I, I address the shift towards method pluralism by discussing pragmatic clinical trials as a novel method to generate evidence, while in the second part, I address the shift towards data pluralism by evaluating emerging practices of data quality assessment for real-world data. My work builds on two in depth case studies. Part I introduces the case of the Salford Lung Study, a pragmatic clinical trial using electronic health record data to investigate the effectiveness of an inhaler to treat COPD. My analysis of this case will focus on the problem of valid causal inference and the question of the practical relevance of evidence. Part II introduces an observational study using registry data to establish the effectiveness of an immunosuppressant for lung transplant recipients. My analysis of the second case will focus on conceptual and methodological issues of data quality assessments, the question of epistemic power of evidence and the problem of trustworthiness of data.

My work on the case studies is supplemented by insights from published medical research articles within this development, from training courses and webinars dedicated to researchers on real-world evidence and pragmatic trials, and from numerous opinion papers by medical professionals and meta-research on the topic. The most invaluable source of materials on which my study builds were the numerous guidelines and white papers published by regulatory agencies, service providers, and expert groups as well as the legal frameworks that govern clinical trials and real-world evidence. Moreover, my research benefited from informal interviews and discussions with practitioners in the field including medical scientists, physicians, and regulators. Finally, I benefited from working alongside experts in the field of clinical research for the past five years in clinical research infrastructures devoted to the facilitation of clinical trials and health data sharing in

Switzerland. I tried to pay justice to the opinions and rationales articulated in these materials while critically studying the risks and opportunities that this evolution will bring. The goal of the study is to critically evaluate the use of real-world data for assessing the effectiveness of medical interventions, so the best can be made out of the challenges and opportunities that this evolution will bring.

Part I:
Method pluralism

Chapter 1
The gold standard

In medical research, randomised controlled trials (RCTs) are widely regarded as the gold standard for drawing reliable causal conclusions. Since the 1970s, the US Food and Drug Administration (FDA) has relied on randomised trials to make decisions about drug approvals, and since the advent of evidence-based medicine (EBM) in the 1990s, the method has become indispensable for informing healthcare decisions. At the heart of the EBM approach is the noble idea that treatment decisions should rely only on high-quality evidence. Its biggest success is the hierarchical theory of evidence with randomised trials ranking near the top of the evidence hierarchies that order methods according to their quality. The hierarchical theory of evidence was so successful that randomised trials nearly replaced all other types of evidence for decision making, such as consensus conferences, mechanistic knowledge, or clinical expertise.

With the advent of real-world evidence, the evidence standard is about to change. Real-world evidence challenges the hierarchical theory of evidence, and with it the dominance of the randomised trial, and instead proposes a contextualised approach to evidence standards. Observational studies are the most prominent and controversial method that has (re)emerged within the real-world evidence paradigm. However, there are other variations of statistical methods for analysing real-world data. They include pragmatic clinical trials, which are randomised trials that are embedded in healthcare settings and can measure treatment effects under natural conditions. In addition, registry-based trials randomise patients from an observational cohort and use the rest of the cohort as the control group. Finally, there are externally controlled trials, in which a single group study is conducted, with a control group that is not part of the same trial. In short, real-world evidence is more than just a new name for observational evidence. However, the randomised method remains the reference standard against which these new methods must be validated and tested.

In the last 20 years, the philosophy of EBM has poked so many holes into the evidence hierarchies and the randomised method that the method's role as the reference standard for evaluating new methods has lost its meaning. Critics have addressed the unjustified downgrading of mechanistic knowledge or clinical expertise (Clarke et al. 2013, 2014) and emphasised the poor external validity of

randomised trials (Cartwright 2007, 2009). They have also argued that quality assessments need to be based on method tokens rather than types (Stegenga 2018) and that randomised trials are ill-suited to demonstrate harmful effects (Osimani 2013). They have also questioned the epistemic pertinence of randomisation itself (Worrall 2007; Worrall 2002). Only a few philosophers have defended the randomised trial as a justified gold standard (Martinez and Teira 2021; Teira 2020).

The goal of this chapter is to clarify some ground rules. The first rule is this: The methodology of the randomised trial is a good reference standard because it can produce unbiased statistical evidence to inform causal claims; furthermore, those who rely on such evidence for decision making can generally distinguish between successful and unsuccessful instances of the method. The rule does not mean that every instance of a randomised trial succeeds in producing unbiased evidence. It also does not mean that a randomised trial is necessary to produce causal knowledge; nor does it mean that unbiased statistical evidence is all we require for our practical and epistemic aims. Most importantly, I do not mean to embrace all properties of the randomised method that are not essential for the production of unbiased evidence. Drawing a distinction between essential and non-essential properties of the method to produce such evidence is a recurring topic in the first part of this thesis.

The second rule that I establish is that there is just one essential property of the randomised method, namely, proper randomisation. 'Proper' comes with several requirements attached. Randomisation is 'proper' only if it is correctly implemented and maintained throughout the trial. This requires not only randomising patients to treatment groups but also concealing the randomisation sequence before randomisation, as well as controlling problems like drop-out rates, cross-over and contamination problems after randomisation. I elaborate on these problems below. The good news for critics of the randomised method is that many of the widely criticised properties of the randomised method are, in my view, non-essential properties. These include homogeneity of the study population and strict eligibility criteria; small sample sizes; placebo controls; short-term outcomes; and, most controversially, blinding. I argue that blinding is not necessary in trials that are in line with patients' preferences, and may be insufficient in trials that are against patients' interests. I hold that aligning trials with patient preferences, if possible, is the better strategy to minimise the risk of bias than increasing control over their behaviour.

If experimentalists introduce some flexibility regarding the non-essential features, the randomised method can deal with a remarkably wider range of

problems than it is currently credited with. Indeed, one of the promises of analysing real-world data within randomised trials is that we can increase the feasibility of large or long-term and heterogeneous trials. Two substantial concerns regarding this promise, however, are the worry that the blinding of patients – the practice of concealing which treatment each patient receives – might be necessary to address problems related to drop-out rates, cross-over and contamination and is therefore necessary for valid causal inference. The second concern is that blinding is equally necessary to deal with the problem of placebo effects for valid causal inference. Although real-world evidence can be generated using randomised procedures, such evidence is generally not obtained under blinded conditions. To respond to these concerns and provide a justification for the two rules, I begin my philosophical inquiry with several old debates about the value of randomisation and blinding in clinical trials.

My work proceeds as follows. In sections 1.1–1.2, I defend the epistemic value of randomisation against Worrall's famous objection of the infinitely many confounders. I show that Worrall's arguments misfire because he neglected the role of statistical significance testing. Section 1.3. defends the claim that randomisation is valuable because its success conditions are accessible to those who rely on the evidence for decision making. Section 2 turns towards issues of blinding and placebo controls. Section 2.1 introduces the methodological rationale of blinding and placebo controls and differentiates between two roles of blinding, the first being to prevent protocol violation and the other to deal with placebo effects. Sections 2.2. and 2.3 discuss two arguments in favour of blinding articulated by Teira and colleagues. The first argument relates to its role for enforcing the protocol (section 2.2) and the second to its role as a warrant for the non-interference condition (section 2.3). I argue that blinding is not necessary to enforce the protocol in trials that are in line with patients' preferences, whereas blinding might be insufficient in trials that are against their interests. The problem of non-interference reinforces the lesson that well-organised patient communities might reinforce the problems of placebo effects and protocol violations, however, I argue that the lack of blinding does not pose an additional problem of non-interference. The last section (section 2.4) proposes additional measures to strengthen the enforcement of proper randomisation and deal with placebo effects.

1. The epistemic value of randomisation

The basic idea of a randomised trial is simple. Researchers select an eligible study population, split the population into two groups (the treatment group receiving the intervention that is being tested, and the control group receiving a control intervention) and randomly assign the patients to these two groups. At the end of the experiment, the two groups are compared with each other regarding an outcome of interest. If there is a clinically relevant and statistically significant difference in outcomes between the two groups, the method allows the researcher to identify the intervention as the cause of the difference. The RCT and the rationale for causal conclusions are often explained in terms of Mill's method of difference. The idea is this: In scenario A, we observe a certain effect E, whereas in scenario B, no effect E is observed. If all but one of the causally relevant factors F_1–F_n are identical in both situations, we eliminate the identical factors as possible causes until only the single remaining cause is left. The elimination of alternative explanations is a deductive inference using *modus tollens*:

1) If F_n is the cause of E and F_n is present in all situations, we can observe the effect E
2) F_n is present in all situations
3) We do not observe the effect

From 1)-3): F_n is not a cause of E

Although the method provides deductive logical rigour, it assumes that the list of alternative hypotheses is complete – and also known to us. Therefore, the method of difference is often classified as an inductive method known as 'eliminative induction.' The precise classification of the method, however, is subject to debate. Pietsch holds that Mill's method of difference is an instance of variational induction, which confirms the causal relevance of circumstances for phenomena rather than eliminating competing hypotheses (Pietsch 2021). Cartwright uses Mill's method of difference to argue that ideal randomised experiments are a deductive method.

Mill's method of difference relies on a comparability or homogeneity assumption between the two situations. That is, all causally relevant factors, except the factor of interest, must not change in value across the scenarios that are being compared. This assumption is highly unlikely to hold for randomised trials. It is a truism that the objects of biomedical research are prone to variability, and no scientist would affirm that it is possible to compare two populations that are exactly alike. Indeed, dealing with such variation is the reason statistical theory has become indispensable and lies at the heart of biomedical research. It is also not highly plausible that scientists have a complete list of causal factors at hand. Randomised experiments are valued precisely because they can license causal inference under the assumption that there are unknown causal factors.

A common attempt to bridge these discrepancies is to think that randomisation can ensure that the homogeneity assumption holds. Broadbent, for example, states that randomised trials are not controlled in the classical sense but rather employ randomisation as a substitute for the classical sense of control (Broadbent 2013, chapter 1). To assess the rigour of RCTs, Broadbent states that the pertinent philosophical question concerns the extent to which randomisation is epistemically equivalent to control in the classical sense. This interpretation is strongly reinforced by a recurring claim in the methodological literature that sounds like the Millean homogeneity assumption. The argument here is that randomisation is sufficient to 'balance all known and unknown confounding factors'. Attempts to draw a link between Mill's homogeneity condition and the role of randomisation, however, seem to have led to serious misunderstandings about the epistemic role of randomisation. A similar point has recently been made by Martinez and Teira, who distinguish between three conceptions of balance. These authors argue that Millean balance is not required in randomised trials, but only what they call 'fisherian balance' and 'efficiency balance' (Martinez and Teira 2021). In a similar vein, Baetu distinguishes between two methods for causal inference in the biomedical sciences, namely directly controlled experiments and randomised experiments and argues that only the former can be explained by using Mill's method of difference (Baetu 2020).

Building mostly on the work of statistician Stephen Senn, I support the notion that the role that randomisation plays in (frequentist) statistical inference can, at best, be a highly sophisticated version of Millean balance. In the next section, I reconstruct the statistical rationale to establish two claims that disentangle the relation between randomisation and homogeneity. First, randomisation has an epistemic role to play for valid causal inference, but it has no bearing on homogeneity in Mill's sense. Second, homogeneity in Mill's sense does play an epistemic role in randomised experiments, but it has no bearing on valid causal inferences.

1.1. The 'infinitely many confounders' objection

Two decades ago, Worrall published an influential critique about the claim that randomisation can balance all known and unknown causal confounders (Worrall 2007; Worrall 2002). His argument is intriguing: Since randomisation is a probabilistic process, it cannot guarantee that causal factors are balanced in any actual trial; instead, they could well be unbalanced by chance. Moreover, Worrall

continues, given that there are plausibly an infinite number of potential unknown confounding factors, it is certain that at least one of them is unbalanced. Hence, the strong claim that randomisation can balance confounding factors applies only in the long run – but this does not help to ensure balance in an actual trial. As a positive proposal, he states that we should instead rely on our knowledge about potential confounders and actively allocate groups such that these confounders are balanced.

The core of Worrall's argument has been well-received by most scholars and is often reiterated to defend the claim that randomised trials are unduly ranked at the top of the evidence hierarchy (Rocca and Anjum 2020; Borgerson 2009; Cartwright 2017; Clarke et al. 2014). For example, in her earlier work, Cartwright built on Worrall's critique and drew a distinction between *ideal* RCTs and *real* RCTs – which, I believe, is partially motivated by the discrepancies between Mill's method and randomised trials. In an ideal RCT, all the theoretical assumptions for a deductive inference are, by definition, met. This includes the homogeneity assumption: 'By definition of an ideal RCT, [the confounding factors] are distributed equally in both the treatment and control wing' (Cartwright 2009, p. 64). In another article, she calls this requirement simply the 'idealization assumption' (Cartwright 2007, p. 16). While her critique on randomised trials is mainly focused on their limited scope of applicability, or their external validity, she also criticises real RCTs for falling short of being ideal RCTs. Among other problems, she emphasises that in real RCTs, randomisation can go wrong because it often does not make it the case that the idealisation assumption is met for an ideal RCT. Cartwright together with Deaton modified this position in a later piece on randomised trials. Nonetheless, she still seems to consider that the requirement for perfect balance of confounding factors is an appropriate starting point to explain and criticise the method (Deaton and Cartwright 2018).[1] Others have incorporated Worrall's critique by qualifying the balance claim with a probability clause. La Caze, for example, writes that randomisation 'improves the probability that

[1] They attempted to reconcile the requirement for balance with the frequentist position that I introduce below. Cartwright and Deaton follow the frequentist position in that they distinguish between the precision and unbiasedness of an RCT result. They seem to acknowledge that a randomised allocation, without balance, only raises concerns for biasedness if (for example) the allocation sequence was not truly random because the random number generator failed (Deaton and Cartwright (2018, p. 14). Their counter argument against randomisation is then directed towards its 'wastefulness' compared to using prior knowledge. This point mirrors the frequentist position regarding the importance of balance or homogeneity for precision rather than unbiasedness (Deaton and Cartwright 2018, pp. 17–18).

extraneous risk factors (known and unknown) are roughly balanced' (La Caze 2017, p. 201). This point is what La Caze calls a 'fragile' claim, which cannot deliver what is implied by many of the unqualified statements about randomisation (La Caze 2017).

Worrall also has critics. Miriam Solomon, for example, acknowledges the validity of the argument, yet claims that Worrall is merely making a logical point with (almost) no practical relevance. According to Solomon, Worrall's concern is only relevant in a case where many unrelated population variables could influence the outcome (Solomon 2015, Chapter 6). Similar arguments have pointed out that it is not the balance of individual causal factors that matters but rather the overall potential outcomes among the participants (Dahly 2019; Fuller 2018). Still others have argued that Worrall's proposed alternatives to randomisation fail to outperform randomisation in terms of balance (Larroulet Philippi 2022). More recently, philosophers followed a line of argument by Worrall's strongest critic, statistician Stephen Senn, and have argued that the whole idea of balance is misguided (Baetu 2020; Martinez and Teira 2021). One does not need to delve into the details of statistical significance testing to see that Worrall's arguments misfire. Baetu proposes a clever argument to make that point. As he points out, the purpose of statistical significance testing is to reject the hypothesis that an observation occurs by chance alone. Yet, if randomisation could **ensure** that all relevant factors are – in Mill's sense – homogeneously distributed, we could simply eliminate the chance hypothesis on **a priori** grounds. The existence of statistical techniques to perform that task is evidence that the role of randomisation has been misidentified (Baetu 2020).

I follow the same route of criticism, but I attempt to restore the meaning of balance in this context by drawing an analogy between random *sampling* and random *allocation*. In short, Worrall too easily dismissed the epistemic significance of significance testing when he wrote:

> I shall not consider this often-examined argument [The Fisherian Argument from the Logic of Significance Testing] in any detail here (it is in any event not the one that has carried most persuasive force sociologically speaking). I just report *first* that it is not in fact clear that the argument is convincing even on its own terms; and *secondly* that there are, of course, many – not all of them convinced Bayesians – who regard the whole of classical significance-testing as having no epistemic validity, and hence who would not be persuaded of the need for randomisation even if it *had* been convincingly shown that the justification for a significance test presupposes randomization. (Worrall 2002, p. 321)

Together with Senn, Martinez and Teira, and contra Worrall, I hold that the practice of significance testing is convincing on its own terms. Hence, I argue that Worrall unsatisfyingly neglected its role for the epistemic value of randomisation. To demonstrate my point, I spend some time reconstructing the logic of statistical significance testing. As I show, by understanding the practice we can disentangle the relation between homogeneity and valid causal inference. Furthermore, disentangling this relation clarifies a distinction between essential and non-essential properties of randomised trials. Some philosophers, of course, would still object that the focus on frequentist statistics is unjustified and that arguments developed within a theory-dependent approach are irrelevant, as such arguments do not apply to those philosophers subscribing to Bayesian methodology. I would respond as follows: first, the vast majority of clinical trials do not follow a Bayesian methodology but employ frequentist statistics. Even guidelines at the highest organisational levels, such as the International Council for Harmonisation (ICH) E9 Guideline on Statistical Principles for Clinical Trials, focus only on frequentist statistical methods (International Council for Harmonisation 1998). Second, I draw attention to carefully cashed out arguments regarding the epistemic superiority of randomisation over the proposed (Bayesian) alternatives (Larroulet Philippi 2022) and (Martinez and Teira 2021) who argue that within Bayesian methodology, randomisation can provide consensus about which covariates to consider if priors about their relevance differ.

1.2. The logic of statistical significance testing

Random processes play a role in two different practices, random *sampling* and random *allocation* in experiments (called random*isation*). These techniques are based on two different theoretical frameworks and should not be confused. However, they do share some high-level features, and I use these features to draw an analogy between the two and elaborate on their epistemic value. I argue that we can understand the notion that 'random allocation balances confounding factors' as analogous to the notion that 'random sampling can ensure representativity'. In both cases, the phrase can be understood in a probabilistic sense that does not fall prey to Worrall's critique when understood in the context of statistical significance testing. I begin by explaining the value of random processes in the framework of random sampling. I then apply these insights to random processes in experiments.

In a world filled with variability, statistical inference is a crucial scientific method to understand what limited observations indicate about an entire population. The rationale to generalise such familiar inferences from a sample to

the population is based on the technique of random *sampling*, where researchers select the individuals to be observed based on a random process. The process of random sampling introduces a probabilistic risk that the observed individuals constitute, by chance, a set of more extreme cases than the other unobserved individuals in the population. However, what is important is that the process of random sampling ensures that every individual in this population has an equal chance to be observed. This assumption can support an objective probability model against which the actual observation can be tested. The job of statistical significance testing is to attach an uncertainty estimate to an observation regarding the risk of a particular observation having occurred by chance alone because – by chance of random sampling – we were looking at extreme cases only.

The practice of significance testing first requires that the scientist needs to stipulate what an unsurprising observation would look like, which is captured in the null hypothesis. The null hypothesis usually states that nothing unusual is expected; hence, all observations are entirely due to variability. For example, there is no difference between an observation and a historically expected value. Under the assumption that each instant had an equal chance to be observed, it is then possible to calculate the probability of obtaining the observed result or a more extreme result. The conclusion of such a statistical test is a conceptually complex proposition. It is a probabilistic proposition that is conditional on the truth of the null hypothesis. The null hypothesis is itself a comparative proposition; it requires comparing a measured value with another value of interest. Typically, significance statements take the following form:

(1) The probability that the observed result occurred by chance alone is α or a smaller value, if the null hypothesis is true.

Scientists reject the null hypothesis if the resulting probability of an observation resulting from chance alone is sufficiently small. This would mean the result is statistically significant.

For random sampling to support such a claim, it is not required that the process of random sampling generates a representative sample in the sense that all the relevant traits occur in identical proportions among both the sample and the population. As we have seen, it is possible that the random process produces an extreme set of observations by chance. What is required, however, is an objective probability model that allows for the calculation of an uncertainty estimate for the observation. The role of random sampling in this procedure is precisely to support this objective probability model by supporting the assumption that all individuals

had an equal chance to be observed in the experiment. Hence, a random sample is considered *representative* of the population simply in virtue of being a *random* sample. In the context of statistical testing, reading this notion to mean that all relevant characteristics are proportionally represented would be an inadequate interpretation. In addition, this is an unnecessary assumption for the statistical inference to be valid.

Similar reasoning can be applied for the case of random *allocation* (or randomisation) in interventional studies but there exists also an important conceptual difference. Random *allocation* refers to allocation of preselected subjects to two or more groups in a comparative experiment, where the allocation and not the selection of subjects is supported by a random process. In a random sampling scenario, researchers draw an inference from the sample to the population from which the sample has been drawn. In a randomised experiment, statistical inference does not extend to the population from which the sample is recruited. Instead, randomisation creates a hypothetical population consisting of *all possible group allocations* of the people enrolled in the study. Within Fisher's framework this allows to distinguish between observed differences between the groups that occurred by chance and differences that occurred for other reasons. Let me elaborate.

In Fisher's framework, causal factors (other than the intervention) that contribute to the outcomes – called covariates – merely contribute to the overall variability of an observed effect. Thus, the more influence from covariates, the more variability can be observed. To distinguish between the causal influence of covariates and the intervention, Fisher's framework compares the variability of the outcome *within* the groups to the variability *between* the two groups (Senn 2013; Martinez and Teira 2021). Under the null hypothesis, the within-group variability and the between-group variability should be equal – but only if all individuals had an equal chance to be observed in either of the two groups. This is where randomisation comes into play. By allocating the individuals to either of the two groups at random, the researcher provides a justification for the assumption that all individuals had equal chance to be observed in either of the two groups. This assumption supports the objective probability model against which the observation can be tested. The valuable epistemic result of this process is that the risk that random allocation may, by chance, generate two extreme groups is indeed considered in the assessment of statistical significance:

> As already explained, conventional analyses of randomised trials make an allowance for the distribution of unmeasured confounders. They do this

> by judging the probability with which the groups can differ from each other by looking at the way in which results differ within groups. Unmeasured confounders make a contribution to both of these measures of variation (between and within group), and the comparison of the two is the cornerstone of the technique of analysis of variance developed by RA Fisher in the 1920s. (Senn 2013, p. 1446)

In other words, the risk of making an error because of unbalanced covariates is already included in the risk of accepting a false-positive or false-negative error that is bound by the statistical significance threshold. No additional risk comes into play here. Consequently, unbalanced covariates are no longer confounders that pose any additional risk for the inference. Instead, they are what frequentists would call a random error that is well-controlled by statistical theory. In short, in the context of significance testing, randomisation can turn the risk of potential confounders into well-controlled statistical error (Senn 2013).

Hence the randomised experiment supports the claim that, within the calculated uncertainty boundaries, the difference would have been observed in all possible group allocations[2]. Among other things, this conception of a hypothetical population in frequentist randomised trials is what Bayesian statisticians take issue with (Howson and Urbach 2006, chapter 6). For Bayesians, it is indeed far from obvious what information about hypothetical populations and possible observations would add to the statistical inference. Yet, for frequentists, such information is precisely what is required for the objective probability model against which the actual observation can be tested – and its uncertainty quantified.

It follows from the above discussion that the epistemic pertinence of random allocation does not rest on the fact that causal factors are 'balanced' or 'distributed equally' in Mill's sense. What is critical is that they are distributed *randomly* (Senn 2020). Indeed, the role of randomisation resists being reduced to ideas about balance or homogeneity, at least not in Mill's sense. More importantly, the idea that randomisation can 'go wrong' or that researchers 'get unlucky' – in any sense other than by manipulating the random process – is clearly mistaken. One might wonder then why the concepts of balance, comparability and homogeneity persist

[2] The frequentist theory is far from providing a universal route to causal inference, mostly because statistical inferences are not thought to be causal in nature. Martinez and Teira assume that causal background assumptions from the experimenter add the causal knowledge that is needed and simply write about statistical and causal inference simultaneously. Baetu on the other hand argues that it is the context of the controlled experiment – i.e., an accurate intervention with a standardized context – that does the heavy work for the causal interpretation of the statistical result.

in the literature. In my view, the analogy with the case of random sampling can help to restore the meaning of these assertions. As I discussed above, random sampling does not ensure that all relevant traits are represented in the sense that all causal factors occur in identical proportions to those of the population. Rather, a sample is representative in a purely statistical sense, namely in virtue of being a random sample. Randomly allocated groups are homogeneous, comparable or balanced in the same way; that is, not despite their randomness but because of it. In both instances, there is an implicit probabilistic reading underpinning their use that does not fall prey to Worrall's critique, if properly understood in the context of statistical significance testing.

So far, I have shown that randomisation is relevant for valid statistical inference within the frequentist framework, although it has no bearing on balance or homogeneity in Mill's sense. Another puzzle that emerges from this discussion is that there is an apparent tension with scientific practice. For all we know, researchers use practices that attempt to increase homogeneity in Mill's sense, for example, by strictly restricting the study population or stratifying the population. The conceptual tool to explain this tension is the distinction between validity and precision in statistical inferences. Homogeneity in Mill's sense can increase the experiment's precision but not its validity (Senn 2013, 1989). A comprehensive reconstruction of Senn's argument has been made by Martinez and Teira, which I do not repeat here (Martinez and Teira 2021). The distinction between precision and validity is explained with the image of a dartboard. If one repeatedly throws darts at the bullseye, each arrow would represent the point estimate of an experiment, whereas the bullseye is the true value. In repeated experiments with low precision, the arrows will be scattered in a wide circle with the bullseye at its centre. In an experiment lacking validity, arrows will be scattered in a circle with a centre that is skewed away from the bullseye. By increasing the homogeneity in Mill's sense, we increase the experiment's precision. This means the arrows will create a smaller circle. Reducing the experimental bias increases its validity, which means that the centre of the circle shifts towards the actual bullseye. A lack of precision is evident in the estimate as a large confidence interval, which indicates the uncertainty boundaries of the estimate. By contrast, a lack of validity is not directly visible.

The distinction can be utilised to explain the view that larger sample sizes make it more likely that causal factors are roughly balanced. It is true that an increase in sample size increases the balance, in the sense that it becomes more probable that important factors will be evenly distributed among the groups.

However, following Senn, this fact has no bearing on the validity of the inference, but only on its precision:

> When sample sizes increase, it is certainly the case that the expected random difference between two groups will reduce, and this reflects, amongst other things, the greater expected balance in proportionate terms between groups. In this sense, the belief that larger trials are more balanced than smaller ones is not a myth. However, by the same token, the standard error of the treatment effect will be smaller and the confidence interval will be narrower, and for any given observed difference at outcome, the p-value will be smaller. Thus, the effect of increasing sample size is *consumed by conventional analyses in terms of increased precision.* There is no further benefit in terms of increased validity. (Senn 2013, p. 1446)

Philosophical frameworks rarely account for the difference between validity and precision. In philosophical terms, these are both epistemic risks. I believe the difference matters epistemically for two reasons. First, the uncertainty that is introduced by a lack of precision is known in precise probabilistic terms, in the form of the confidence interval or the p-value; this fact allows fine-grained conventions to be formed about the amount of risk the scientific community is willing to take. The same does not hold for validity. The severity of this epistemic risk lies precisely in the fact that it remains hidden in the statistical quantification of uncertainty. Instead, it needs to be made visible with measures such as risk-of-bias assessments (see Chapter 2). Consequently, in scientific practice, there is officially zero tolerance towards a lack of validity. Second, precision can be controlled by several parameters that can be tailored to each other to ensure a fixed and deemed acceptable level of uncertainty. Thus, wide variability within the trial can be compensated by a larger study population to achieve sufficient precision; and small sample sizes can result in sufficient precision in trials with high homogeneity. For these two reasons, I suggest keeping precision and validity apart. Certainly, we require both of them to make a valid causal inference with acceptable uncertainty. However, because precision is not subject to a zero-tolerance norm and can be achieved by several complementary properties, we can consider each of the properties that pay into the inference's precision as non-essential for valid and informative causal inference. This is relevant because randomised trials have been widely criticised for using properties like the many exclusion criteria or highly standardised training processes because this minimises the method's relevance for patients outside the scope of the trial (see for example Travers et al. 2007). If randomised trials do not require any of these properties to provide valid causal inference, this is good news.

1.3.Accessibility of success conditions

In the previous section 1.2, I argued that randomisation together with the statistical machinery of significance testing can support valid causal inference. Whether randomised trials are better than other statistical methods – or just equally good as them – in supporting valid causal inference remains to be discussed. The superiority of randomisation has been defended on various grounds. Even Worrall accepts that randomisation is a good (although unnecessary) mean to prevent the problem of selection bias, which is the problem that human judgement can – intentionally or unintentionally – distort the allocation procedure (Worrall 2002). Senn has argued that randomisation is a necessary means of maintaining blinding and preventing manipulation (Senn 2013, 1994). Larroulet Philippi holds that randomisation purports rational stability and rational agreement (Larroulet Philippi 2022) and Martinez and Teira argue that it is a good mean for resolving disagreements (Martinez and Teira 2021). Furthermore, randomisation is often seen as a fair procedure to allocate subjects to unequal interventions (Kombe et al. 2019) .

Another epistemic advantage of randomisation that I want to highlight is the simplicity of the mechanical procedure. The successful design of non-randomised studies relies on causal background knowledge and expert judgements, including knowledge about relevant covariates and the choice of appropriate methods to control for their influence, as well as skills in applying those methods successfully. Hence, for non-randomised studies to be successful, our knowledge of confounding factors must be true and complete, and researchers must be sufficiently skilled and impartial. These conditions are not only hard to satisfy but also difficult to verify. Consequently, it is not uncommon to read statements like the following, which appears in a guidance paper by the Swiss regulatory agency Swissmedic: '[S]tatistical methods to adjust for, e.g., unbalanced baseline characteristics often rely on *subjective* assumptions with respect to the relevant factors' (Swissmedic 2022, p.2, my emphasis). While prior knowledge is not subjective in the strong sense of the term, determining what prior knowledge is relevant to consider is prone to error and disagreement. Furthermore, the assumption that prior knowledge is complete is not epistemically accessible.

Randomisation replaces these assumptions with a simple mechanical procedure. This procedure is robust under the assumption that prior knowledge is incomplete or incorrect, and it is also robust under the assumption that researchers make mistakes or try to manipulate results in their favour. The epistemic

advantages of this mechanical procedure are not only that randomisation is less fallible than relying on skilful use of prior knowledge, but also that its success conditions are accessible to verification. Indeed, even external parties who do not have the same level of medical and scientific expertise as the researchers can verify the success conditions of randomisation. All that is required is that researchers follow a simple mechanical procedure. In Chapter 6, I expand on this idea. I argue that oversight by external parties, such as regulators or ethics committees, is essential to ensure the trustworthiness of clinical evidence. For these external parties to play their role effectively and judge whether an experiment is reliable, the conditions for success of an experiment must be accessible to them.

What the research community came to appraise as a great advantage has been criticised by Cartwright and Deaton, who stated that 'The systematic refusal to use prior knowledge and the associated preference for RCTs are recipes for preventing cumulative scientific progress' (Deaton and Cartwright 2018, p. 19). I do not share their intuitions about scientific progress or the intuition that a preference for randomised trials amounts to a 'systematic refusal' of prior knowledge. I put these issues aside for now. In this chapter, it is irrelevant how randomised trials relate to a particular theory of scientific progress; what matters is that randomised trials, in a given instance, provide epistemic advantages over the use of prior knowledge. They are reliable even if the prior knowledge is erroneous; they are feasible even if prior knowledge is incomplete; and they come with epistemically accessible success conditions that allow to distinguish between successful and unsuccessful instances of the method. In the context of medical and pharmaceutical research, the reliability and trustworthiness of a particular causal conclusion matter more to patients than does long-term scientific progress.

In summary, I have argued that randomisation – together with the statistical machinery of significance testing – can support valid causal inference. Hence, Worrall's arguments fall short because he neglects the important epistemic role of the practice of significance testing. It is within this context that we can make sense of the ubiquitous talk in the literature about 'balance' of experimental groups namely in the same way as we can make sense of the 'representativity' of random samples. The good news for the critics of the method is that increasing the homogeneity of samples in Mill's sense, although a common practice, is not essential for the validity of the causal inference. Finally, I have argued that randomisation comes with the epistemic advantage that its success conditions are easily accessible for verification by external parties.

2. The epistemic value of blinding and placebo controls

Blinding and placebo controls are techniques that have emerged specifically for experimental purposes. Blinding refers to techniques that withhold information about treatment allocation from patients – and sometimes also care givers and outcome assessors. A placebo control is a dummy treatment that mimics the test treatment in its phenomenological properties but has no effect on the target disease. Data from routine data collection processes is thus never obtained under blinded or placebo-controlled conditions. This point has already raised major concerns among critics of the real-world evidence movement. Fraile Navarro, Tempini and Teira argued that pragmatic trials fail to provide informative evidence about the effectiveness of treatments, mostly on the ground that these trials are unblinded (Fraile Navarro et al. 2021). Hemkens pointed out that 'a specific beauty' of real-world outcomes data is that the data are formally blinded, because people who collect such data are generally independent from a research endeavour (Hemkens 2018). The epistemic value of blinding has been a matter of debate for clinical trials in general. Howick cast doubt on the assumption that blinding is necessary to maintain the evidential rigour of the gold standard (Howick 2011, Chapter 6), whereas Teira argued that blinding is necessary for the integrity of the causal inference as a warrant of the non-interference condition (Teira 2013).

In my contribution to this debate, I argue that blinding is not necessary in trials that are in line with patient preferences. However, in trials that go against the interests of patients, blinding is necessary but may not be sufficient. Consequently, aligning trials with patient preferences, if possible, is the better strategy to minimise the risk of bias than increasing control over their behaviour. Chapter 2 expands on this idea and shows how pragmatic clinical trials can do without blinding because these trials serve a different purpose.

2.1. Problems with the methodological rationale

Blinding and placebo controls are intimately intertwined. It is common but too simplified to conceptualise a placebo control as an 'inactive treatment' with 'no effect'. Placebos often do have effects and are sometimes designed to replicate common side-effects of the test treatment. To clarify the concept, Grünbaum introduced a distinction between 'characteristic features' and 'incidental features' of a treatment. Characteristic features are related to treatment of the condition by

a therapeutic theory (e.g., the mechanism of action), while all else is incidental. Placebos for a certain disease then replicate the incidental features but none of the characteristic features (Howick 2017). Methodologists particularly encourage the use of placebos in trials where no alternative treatment exists; however, active-controlled trials can also use placebos to maintain the blinding.

Trials use placebo controls for three reasons. The first is that such a trial allows the researcher to subtract the effect from the incidental features of a therapy by replicating all the incidental features and their effects in the control group. The net effect of these additional components (together with the effect of expectations) is usually subsumed under the 'placebo effect' of an active treatment. The remaining effect is thought to be the *absolute* or *true* effect of the treatment (Howick 2011, Chapter 8). Interestingly, placebo controls have become entrenched into notions of effectiveness, as treatments are usually considered effective if they are more effective than a placebo in a well-controlled randomised trial. A second reason for using placebos is that they make the blinding of patients and physicians possible by making the placebos resemble the appearance of the investigational treatment. Sometimes, placebo controls even replicate common side-effects of the treatment to enhance the resemblance with the treatment. Here, placebos are merely the physical basis to make blinding possible. The third reason for using placebo controls is to warrant a property of trials called 'assay sensitivity'. A clinical trial is said to have assay sensitivity if it can distinguish between an effective treatment versus a less effective or ineffective treatment. Active-controlled trials can run into problems with assay sensitivity in so-called non-inferiority designs, which only need to demonstrate that a test treatment is not inferior to a control treatment. A suggested solution is to require a demonstration of the superiority of treatments over their control rather than non-inferiority (Howick 2011, Chapter 8). In Chapter 3, I propose a similar argument that placebo controls are helpful properties in ideal trials to increase the likelihood of identifying an ineffective treatment. Therefore, I do not consider the argument about assay sensitivity here. While placebos can help to make certain aspects of treatments visible, they are not required to license valid causal inference and hence non-essential properties.

Blinding refers to techniques that keep patients in a state of ignorance about the treatment they receive. The counterpart of blinded trials is open-label trials, where patients are randomly assigned to a treatment, and their treatment assignment is revealed to them after the randomisation process. Blinding helps to avoid two types of problems. The first problem is another type of placebo effect

that occurs because of people's expectations or beliefs about treatment assignment. By keeping participants in a state of ignorance, researchers avoid the possibility that strong attitudes towards the treatment can be formed and may impact the outcome. Hence, blinding is supposed to break any systematic connection between mental states about the treatment and its success. The second problem addressed by blinding concerns actions that result from patients' awareness about the treatment allocation, which would harm the rules of the experiment and hence undermine its capacity to support a valid causal conclusion. Generally, these rules aim to minimise the number of patients who drop out of the trial and prohibit patients from switching to another treatment group or seeking treatments outside the trial, because such behaviour could bias the treatment effect estimate. These rules of the experiment are defined in the trial protocol; hence I follow Fraile Navarro and colleagues in calling these problems 'protocol violations'. The idea is that if patients are ignorant about what treatment they receive, they have no ground to act upon their interests and violate the protocol.

Sometimes 'allocation concealment' is incorrectly added to the list of techniques covered by blinding. However, it is important to keep these techniques apart (Schulz et al. 2002). Allocation concealment refers to the process of concealing the randomisation sequence. That is, researchers who recruit patients for participation in a trial should themselves be ignorant of whether the next random number assigns a patient to the treatment or the control group. If such concealment is violated, the trial is likely to suffer from bias, because recruiters can then select the patients they think are suitable for the upcoming allocation. This would mean randomisation is overridden. According to Schulz and colleagues, a participant is successfully blinded if both the allocation sequence and the treatment are concealed. This notion is correct; we cannot plausibly argue that blinding was maintained if the allocation sequence was not properly concealed. Yet, it does not follow that a trial cannot – or should not – conceal the randomisation sequence just because it is open-label. Open-label trials can and should conceal the allocation sequence *before* the recruitment to avoid such bias. However, the treatment status can be revealed *after* randomisation has allocated the patients to the treatments. The role of blinding for protocol enforcement is to ensure that patients stay in the group *after* the randomisation, whereas the role of allocation concealment is to ensure that recruiters cannot bias the allocation *before* randomisation. Together, these techniques are what I call *proper* randomisation. Allocation concealment is commonly recognised as relevant for proper randomisation and hence is an

essential feature (Schulz et al. 2002). What remains to be determined is whether blinding is too.

In sum, blinding participants is an attempt to keep them in a state of ignorance about their treatment assignment with a twofold purpose: first, to inhibit formation of attitudes that impact treatment outcome, and second, to further inhibit participants to act upon their interests. Together with the use of placebos, such trials provide not only an *unbiased* estimate of the treatment effect but also an *absolute* or *true* treatment effect that is cleansed from different types of undesired contingent effects (sometimes all subsumed under the placebo effect): Effects from the incidental features of the therapy, effects from attitudes of patients about the treatment and patients' actions resulting from their attitudes about the treatment. In my view we should care a lot about unbiased effects that are ensured by proper randomisation, but we need not necessarily care about true effects that require blinding and placebo-controls.

To begin with, I think the methodological rationale is problematic for a couple of reasons. First, we should note how demanding the requirements are for these techniques to be successful. Howick argues that only *legitimate* placebos can fill this role – those that control for all (and only) the incidental features of a therapy. Howick provides examples of why placebo controls often violate these two conditions and thereby under- or overcontrol the effect of the treatment (Howick 2011, Chapter 7). The same holds for blinding, which is known to be hard to achieve. Second, researchers still have a rather poor empirical and conceptual understanding of placebo effects and the driving factors in patients' behaviour. I elaborate this point with reference to the debate on how to measure successful blinding (Teira 2013; Sackett 2007).

The most common method to measure successful blinding is to ask trial participants at the end of the trial to guess in which group they were placed; blinding is considered successful only if the rate of correct guesses is no higher than chance. This approach appears plausible. However, blinding first and foremost prevents patients from *knowing* in which treatment group they were placed. Hence, the approach is contested for two reasons. First, unblinding that occurs because of a positive effect from the treatment poses no problem for causal inference (Senn 1994). Otherwise, we would run into Phillip's paradox, namely, the problem that the most effective treatments could not be supported by the best evidence, as it is almost impossible to keep people blinded if a treatment is highly effective (Howick 2011, Chapter 6). Second, it is unclear whether *true* beliefs about treatment assignment are really what cause bias. For example, participants who falsely believe

that they receive a placebo could also confound the trial results by dropping out of the trial. A second approach to measuring the success of blinding instead suggests testing for the *persistence* of beliefs, whether true or false. However, this approach also suffers from problems. The first is that it is unclear why blinding should prevent persistent beliefs. Blinding can prevent patients from learning which treatment they receive via an information leak. It cannot, however, stop them from *forming* beliefs, whether true or false and persistent or not. Relatedly, it is not clear whether beliefs about treatment assignments cause the placebo effect – rather than it being caused by preferences, expectations or hopes or something else. Relying on blinding to prevent all of these potentially relevant attitudes seems overly optimistic. The second problem of the persistent belief approach is that persistent beliefs about the treatment do not seem a good indicator for protocol violations, such as patients seeking other treatments. Indeed, as the next section 2.2 illustrates, being in a state of ignorance is a good reason to act upon such ignorance and break the trial protocol. Similarly, some scholars have argued that being in a state of ignorance makes it equally plausible that participants understate the treatment responses. This can happen because participants bias their responses towards a moderate response because of the desire not to err too greatly. The standard interpretation, on the other hand, is that open-label trials overestimate the treatment responses, without further grounds to justify such an interpretation. Hence, as Teira noted, the different approaches to measure blinding imply different conceptions about confounding through unsuccessful blinding (Teira 2013). I concur that these are severe conceptual confusions about what it means for blinding to be successful. One might doubt whether it is sufficient to ground an epistemic necessity claim for blinding in clinical trials if we cannot justify which kind of beliefs might cause bias; why beliefs should be privileged over hunches, hopes or expectations as causes of bias; how blinding could block all of these potentially relevant attitudes; or why ignorance itself cannot cause such bias.

Although I am sceptical about how convincing the methodological rationale is, my main argument does not rely on its plausibility or implausibility. What is relevant for my purposes is that we clearly understand what blinding is thought to deliver, keep blinding apart from allocation concealment and distinguish between the two functions of blinding. One is to control for patients' attitudes to eliminate placebo effects, and the other controls for patients' behaviour to avoid protocol violations.

2.2. Blinding to prevent protocol violations

Clinical trial protocols pre-specify the experimental plan. Their main role is to define the rules of the experiment and coordinate its execution. Since clinical trials are commonly distributed experiments across several hospital facilities, protocols are indispensable to ensure the integrity of the data collection and preserve the unbiasedness of the experiment. Fraile Navarro, Tempini and Teira pointed out that it is not only investigators and study nurses who need to play by the rules – but also patients (Fraile Navarro et al. 2021). The two most important rules for participating patients are that they are not allowed to switch to the other treatment group, nor should they seek additional treatments. Moreover, patients are expected to adhere to treatments as specified in the protocol; that is, they should take the treatment with the frequency and dosage foreseen in the protocol. They are also expected to stay in the trial throughout, although for ethical reasons they are allowed to end their participation at any time. However, playing by the rules is not always in the patients' best interests. Therefore, blinding is used to prevent at least the first two problems. If patients are ignorant about their treatment assignment, they cannot know which action would be in their best interest; hence, their best bet is simply to comply with the protocol.

Fraile Navarro et al. discuss the role of blinding as a mean to enforce the protocol against patients' interests within the context of organised patient communities. What they have in mind are cases of well-organised patient communities that cooperate and even conspire against experimenters to enforce their own interests (Fraile Navarro et al. 2021). They illustrate the problem with two historical cases. Their first example is the famous case of trials on AZT for treating HIV/AIDS. The first phase II trial on AZT was terminated early because evidence suggested significant benefits of the treatment over a placebo. At the time the trial was stopped, 19 participants in the placebo group had died, but only one patient in the treatment group had (Fischl et al. 1987). This strong evidence, together with patients' unwillingness to participate in placebo-controlled trials, spurred vivid discussions about the ethicality and practicality of another well-controlled phase III clinical trial. However, even before the phase II trial ended, rumours emerged that the HIV/AIDS community had broken the trial protocol by analysing the medicines in the laboratory or pooling pills among the participants to ensure that everyone would receive at least some of the active treatment (Epstein

1996)[3]. The second example is from a phase II trial testing a treatment for amyotrophic lateral sclerosis. Trial participants took advantage of the online platform called PatientsLikeMe, which is designed for patients to share their health data and experiences. The participants used the online platform to share their experiences from the trial and eventually successfully unblinded the trial.

The examples are well-chosen to illustrate the authors' claim that patients can successfully defend their interests against the experimenters, to a point where they might even sabotage the trial's protocol. In both cases, patients seriously interfered with the validity of the trial. Hence, the authors are right to guard against harmful interests that are not limited to pharmaceutical companies but also affect patient communities. The examples by Fraile Navarro and colleagues illustrate that in some cases interests of patients can be so strong that blinding is *insufficient* to control patients' behaviour. Indeed, in the HIV/AIDS example, patients pooled their treatments precisely because they were ignorant about the treatment allocation[4]. Hence clinical trials can suffer from a risk of bias despite the researcher's best efforts to control patients' behaviour by keeping them ignorant about their treatment. It is plausible that in this case a trial with less controls for patients' preferences would have suffered from even more biases. However, they generalise this example to an extent that is unsubstantiated. As they state:

> 'In these two vignettes we see the trade-off between impartiality and trial participants' freedom in RCTs at work: the less control on the patients' preferences, the more biases will affect the trial outcome' (Fraile Navarro et al. 2021, p. 7).

What is characteristic of the HIV/AIDS example is that patients have strong interests that conflict with the rules of the trial protocol. In cases where the protocol is in line with patients' preferences, concerns about patients breaking the protocol are less relevant. To illustrate, consider the difference between active-controlled and placebo-controlled trials. Without evidence, a placebo control might be equivalent to an active treatment, from a population perspective. From the patient's perspective, however, a placebo cannot be equivalent to an active treatment. A placebo treatment has, by definition, no positive effect on the condition, whereas an active treatment has at least the potential to have a positive effect until proven otherwise. From the individual patient's perspective, the patient can only lose if

[3] He further refers to the (fictional) work by Lapierre, *Beyond Love*, 366-367.

[4] The HIV/AIDS trial example has been used to support the claim that well-controlled randomised trials are not necessary for conclusive evidence about the causal efficacy of a treatment Solomon (2015).

they agree to receive a placebo treatment; hence, it seems irrational to agree to a placebo if one's health is at stake.[5] In other words, placebo controls commonly run strongly counter to the interests of patients. This point reinforces the view that controlling for patients' behaviour can be crucial to the integrity of a placebo-controlled trial. If we tell patients they are receiving a placebo treatment, we might not expect them to comply with the protocol; such compliance could be irrational. Yet, my point also questions whether we can simply generalise empirical evidence from unblinded placebo-controlled trials to open-label trials that use active treatment or even the standard of care. In an earlier paper, Teira partially addresses this point when he writes as follows:

> People can break protocol if this is in their interests. If the outcome of a trial, be it a drug or a social policy, is important enough for the participants, we cannot expect them to comply with a randomized protocol, *unless they believe the treatments in both arms are equivalent.* (Teira 2013, p. 362, my emphasis)

If patients believe that the treatments in both arms are equivalent, patients have no interest in breaking the rules of the protocol. Hence controlling their behaviour through blinding is unnecessary. It is plausible that an unblinded trial where patients believe that treatment groups are equivalent is subject to fewer biases than a blinded trial that does not fulfil the equivalence requirement. As the historical cases of Fraile Navarro and colleagues illustrate, when patient interests are strong, blinding is insufficient to prevent patients from breaking the trial protocol. One option, and sometimes all that scientists can do, is to increase control over patients when their preferences differ from the protocol requirements. The better option, if possible, is to reduce the tension between patients' preferences and protocol requirements.

Assuming that the standard of care is an effective treatment, trials using the standard of care as the comparator are plausible candidates to fulfil the equivalency requirement. In reality, patients' motivation and attitudes towards participation and compliance with trial protocols are not only determined by their perception of the equivalence of the treatment groups. Their willingness to comply with the protocol is plausibly also related to the severity and natural course of their disease, available alternative treatments and the burden to participate in the trial. For example, if patients are not suffering from a rapidly progressing disease and trust

[5] The sceptic might object that placebos can be a rational option, if one wants to avoid harm. I hold that in such a case the really rational patients would simply not participate in the trial.

that they will receive the investigational treatment if the trial is successfully completed, it is in their best interest to comply with the protocol despite believing that the treatments are not equivalent[6]. In non-severe diseases, widespread altruistic motives to contribute to science might be sufficient to comply with the protocol, despite knowing their treatment assignment. Hence, equivalence between treatment options is only one way to align the protocol with the interests of patients.

2.3. Blinding as a warrant for the non-interference condition

A second argument assigns blinding the crucial role as a warrant for the so-called 'non-interference assumption', which is essential for causal inference in experiments (Teira 2013). If the argument is convincing, it not only states that unblinded trials are problematic because of protocol violations but also that unblinded trials can harm the basic principles of causal inference. The non-interference assumption generally excludes 'spill over' effects. Teira follows Gerber and Green (Gerber and Green 2012), who describe the non-interference assumption more precisely:

> [F]or each participant in the experiment, the value of the potential outcome depends only upon whether or not she or he gets the treatment. More precisely, the potential outcomes that would arise if a subject were affected by the treatment of other subjects are declared negligible. (Teira 2013, p. 359)

To illustrate the assumption, Gerber and Green cite the example of the causal impact of female policy makers on the sanitary budget of their village. The allocation of a female instead of male council is the policy intervention that is tested in seven villages. Assuming non-interference means whatever happens in one village does not spill over or interfere with the outcome in any of the other villages. Accordingly, the assumption is violated if the budget from village A depends on whether village B has a female councillor or not. There are various ways in which we can imagine this to happen. Village B might set an example and inspire village A to spend more; or the opposite: village B might provide a free-ride opportunity and encourage village A to spend less. Whether or not the non-interference assumption is justified in each case, Gerber and Green argue, depends on various local facts and is not easily determined. The geographical relation, the participants'

[6] The sceptic might argue that this scenario is also the best option because patients in the placebo group can avoid harm of an active but potentially ineffective treatment while only receiving the treatment after its effectiveness has been proven.

ways of communication and potential budget dependencies all matter.[7] In other words, non-interference would mean that each village has only two potential outcomes: one if assigned to the treatment group and the other if assigned to the control group. If interference occurs, each village's potential outcomes are multiplied manyfold for each possible assignment of all the other villages. Crucially, non-interference is a necessary requirement for causal inference in the sense that if interference takes place between the two groups, naïve comparison of the mean outcomes is meaningless.

It is mostly in the social sciences where this assumption poses important problems. Other examples discussed by Gerber and Green concern communication interventions and displacement interventions. They also discuss the more relevant example of the chance that a vaccinated person could contract a disease in a way that depends on the vaccination status of others nearby. In clinical trials testing drugs, by contrast, the problem of interference is usually deemed negligible. Teira argues that this assumption is only warranted in blinded trials:

> I am taking pill A, there is no physical mechanism by which it can have an effect on your intake of pill B. This interference can only take place through the expectations of the patients about each treatment: if I think that the experimental treatment is better than the standard alternative and I believe I am receiving this latter, these expectations may impinge on the outcome. There is wide evidence about such *placebo effects,* and in order to prevent them, clinical trials are *double blinded,* if the therapy allows it: the treatments are masked so that participants remain ignorant of which one they are receiving, at least during the initial stages of the trial. In other words, *the masking of treatments constitutes the methodological warrant of the NIA* [non-interference assumption]. (Teira 2013, p. 359)

According to the argument, interference between patients can occur through placebo effects. Certainly, placebo effects are widely acknowledged and have been empirically demonstrated. Yet, the placebo effect is usually conceived like this: It is *my* assignment status, prompting *my* expectations about the effectiveness of the treatment, that will influence *my* outcome, but my expectations will not influence the outcome of other participants. Hence, it is obscure how the mechanism could allegedly spill over to other participants. Non-interference requires that my outcome is independent of the assignment status of *others*, which seems to be the case despite any placebo effects.

In defence of the argument, we might suggest a scenario like the one discussed above. If Peter is part of a well-organised patient community and shares

[7] For an extended discussion of the example, see Gerber and Green p. 43-44 and 253-256.

with other participants in the trial that he experiences certain improvements while receiving treatment A, this news could reinforce the placebo effect of people in the same group. In turn, they might experience similar improvements, which they otherwise might not have experienced. Such effects, however, seem to be no additional problem to the general problem that placebo effects can impact treatment outcomes. Whether the placebo effect has been caused by what I believe about the treatment or by what I hear that others experience from taking the treatment seems irrelevant. If that is the case, however, comparison of mean outcomes is not meaningless; it merely means that we compare the treatments effects *including* any placebo effects. The relevance of the assumption increases for behavioural interventions, psychopharmacological interventions or infectious disease interventions if the people who participate in the trial are part of the same community.

Even if we consider placebo effects problematic for the non-interference assumption, the threat is greater in blinded trials than in open-label trials. In the former, effects more easily spill over between control and treatment groups. If I hear Peter's report about treatment regimen A, while knowing that I am on treatment regimen B, it is less plausible that Peter's report has an impact on my outcome than if I were ignorant about our treatment assignments. My hope, expectation, or hunch that I might be receiving the same treatment as Peter could be absolutely sufficient to prompt a placebo effect – even if I am not in fact receiving the same treatment. By contrast, hopes, hunches and expectations are less likely to influence my placebo response if I know for certain that I do not receive the same treatment as Peter. Without our ful understanding how the mechanisms of placebo effects work, I acknowledge that such arguments are speculative and insufficient to settle the issue. Overall, the non-interference problem in clinical trials is negligible in the sense that it does not pose an additional problem to the problem of placebo effects. In unblinded trials comparisons of the mean outcomes are not meaningless; it is simply that mean outcomes would include placebo effects – which is a different problem.

Communication between participants could be more problematic if it affects the behaviour of patients that violates the trial protocol. If Peter shares with other participants in the trial that he experiences certain improvements while receiving treatment A, this news could prompt other participants to seek the same treatment as Peter receives. In this case, open-label trials have a clear disadvantage. If I hear Peter's report about treatment regimen A, while knowing that I am on treatment regimen B, I have a direct reason to seek the same treatment as Peter. However,

even if Peter and I are both ignorant about each other's treatment assignments, Peter's report might suffice as a reason to switch to the *other* group than the one I am currently allocated if I am not experiencing similar success with the treatment I receive. Being ignorant about the treatment allocation is again insufficient to prevent patients from acting upon their interests. I acknowledge that communication among participants can reinforce the problem that patients act upon their interests and break the trial protocol. I also acknowledge that this problem is slightly more plausible in open-label trials than it is in blinded trials. However, the non-interference problem, again, does not pose an additional problem to the problem of protocol violations.

Overall, if patients are well-organised and communicate with each other about their outcomes, this can clearly reinforce the problem of placebo effects and protocol violations. Both issues can be problematic, but they are not problematic for the non-interference condition. More importantly, the problem that organised communication poses for the validity of clinical trials is not significantly different for blinded and unblinded trials.

2.4. Measures to reinforce compliance with the protocol

I have argued that blinding is not in all circumstances effective or necessary to prevent protocol violations. However, the fact that blinding is sometimes insufficient does not give rise to an argument to dispense with it. Rather, we should complement trials with strategies that reinforce our epistemic aims. Here are two suggestions to that end. First, to minimise protocol violations, the straightforward solution is to align the trial protocol with patient preferences. This could mean that patients receive the standard of care or any other comparative treatment that patients think is equivalent. Contra Fraile Navarro et al., who argue that there is a trade-off between the freedom of patients and the impartiality of the evidence, I hold that including patient preferences does not negatively impact the impartiality of the trial; on the contrary, it helps prevent biases if it motivates patients to play by the rules of the protocol. However, I do not mean to argue that placebo controls should be avoided under all circumstances. The Chapter 3 develops a position about when we should not.

Second, a simple way to reinforce control over protocol violations is by directly testing for such harmful behaviour, like patients dropping out from the trial or seeking additional treatments outside the study. Such strategies have been suggested already (Sackett 2007; Howick 2011). This strategy can measure protocol

violations in both, blinded and open-label trials, and even facilitate a better comparison between the two on a meta-research level. The approach does not minimise the risk of protocol violations but helps to distinguish between successful and unsuccessful instances of a trial. The methodological literature commonly distinguishes two options to deal with protocol violations. Scientists can choose to exclude these patients from the final analysis, that is, only patients who receive and adhere to the treatment as defined by the protocol are included in the analysis. The methodological literature calls this a 'per-protocol' analysis. The epistemic costs of this approach are that it undermines the randomisation scheme. The alternative is an 'intention-to-treat' approach, where all patients are included in the final analysis according to the group they were first assigned to. This approach maintains randomisation but prompts a different interpretation of the estimated treatment effect (Tripepi et al. 2020). Both options are generally considered acceptable, which implies that scientists have some tolerance for protocol violations in real clinical trials. However, it is rarely the case that scientists comprehensively measure protocol violations and (pre-)define limits to the amount of protocol violation that is acceptable. If protocol violations exceed this threshold, scientists should choose a third option: Declare their experiment unsuccessful.

I have argued that blinding is not essential to ensure adherence to the protocol and hence ensure proper randomisation. However, none of the measures discussed so far can address the issue of placebo effects. In my view, placebo effects are problematic if scientists are interested in the *true* or *absolute* treatment effect, but they are not problematic if scientists are only interested in an *unbiased* effect estimate. While unbiased treatment effects are essential, the importance of true treatment effects depends on the purpose of an experiment. The standard distinction between the 'per-protocol' analysis and the 'intention-to-treat' analysis already illustrates that the concept of a true treatment effect can be interpreted differently. I expand on this idea in the next chapter when I discuss blinding and treatment effects in pragmatic clinical trials.

Chapter 2
The epistemic rigour of pragmatic clinical trials

Renowned medical statistician Douglas G. Altman opened his editorial in the British Medical Journal (BMJ) in 1994 by saying that 'We need less research, better research and research for the right reasons' (Altman 1994). His provocative point was that biomedical research is a scandal. There are too many publications that misinterpret results, selectively report them or are built entirely on false premises. Three decades have passed since Altman's editorial but it seems that little has changed. In 2000, Balas and Boren estimated that only 14% of the available evidence is successfully implemented into clinical practice – and this takes, on average, 17 years (Balas and Boren 2000). In 2009, the Lancet series 'Increasing Value and Reducing Waste' estimated that around 85% of the investment in biomedical research is 'research waste', meaning inadequately produced or reported evidence (Chalmers and Glasziou 2009). In 2014, the former editor of the BMJ reiterated Altman's call in his opinion paper 'Medical research – still a scandal' (Smith 2014). In 2016, Ioannidis wrote a piece called 'Why Most Clinical Research is Not Useful' (Ioannidis 2016). In recent years, others have followed suit, estimating that 56% of patients still participate in 'bad' clinical trials (Pirosca et al. 2022) and that only about 26% of randomised trials are informative for clinical practice (Hutchinson et al. 2022). Philosophers have likewise reached rather pessimistic conclusions about the state of medical research and have argued for fewer clinical trials (Borgerson 2016) or have adopted medical nihilism (Stegenga 2018).

The root causes of these problems are manifold. However, one widely acknowledged and highly criticised shortcoming of randomised clinical trials is the problem that their results do not easily apply to routine care and that they lack *external validity* (Cartwright 2009; Rothwell 2005; Borgerson 2013). Indeed, effects demonstrated in traditional trials often diminish or disappear when treatments are introduced into clinical practice. This problem is known as the *efficacy–effectiveness gap* (Eichler et al. 2011). Concerns regarding the external validity of highly controlled trials have commonly supported arguments in favour of observational studies. Another more recently popularised response to the problem

is the so-called pragmatic clinical trial. Pragmatic clinical trials are randomised trials that are conducted under practical conditions and lack many of the control measures of traditional clinical trials. Their promise is that they preserve the epistemic rigour of randomisation while producing practically useful and widely applicable evidence (Borgerson 2013; Schwartz and Lellouch 2009; Thorpe et al. 2009; Ioannidis 2016; Tunis et al. 2003; Mc Cord et al. 2018; Hemkens 2018). Hence, pragmatic clinical trials have been suggested as a potential solution to the vexing problem of extrapolation (Fuller 2019; Howick et al. 2013a; La Caze 2017). Advocates of pragmatism have argued that the research community ought to prioritise pragmatic trials (Zwarenstein and Treweek 2009; Borgerson 2013). Indeed, pragmatic trials are becoming increasingly popular, and some scholars have even proclaimed the 'rise of pragmatism' (Patsopoulos 2011). Within the Food and Drug Administration's (FDA) real-world evidence framework and the ongoing evolution of evidence standards, pragmatic clinical trials play a prominent role because these study designs often allow for outcome data to be collected from real-world data sources (US Food and Drug Administration 2018).

There is no consensus yet among researchers on whether pragmatic trials adhere to the fundamental standards for quality. Indeed, pragmatic trials often violate some of the standard principles for high-quality clinical trials. For example, they often do not blind patients; nor do they control adherence to treatment or treatment delivery practices, factors that are commonly thought to be important for the *unbiasedness* of experiments. Consequently, researchers often say that pragmatic trials have decreased *internal validity.* Fraile Navarro and colleagues explicitly criticised pragmatic trials for their lack of impartiality (Fraile Navarro et al. 2021). Advocates of pragmatic trials – whom I call 'the pragmatists' in the rest of this chapter – counter the critics by referring to the randomised nature of the pragmatic experiment. They also stress that pragmatic trials have a different purpose (Zuidgeest et al. 2017) or they emphasise the practical value of the evidence (Borgerson 2013). The pragmatists' position echoes the newly revised version of an international reference guideline for clinical trial designs by the ICH, which explicitly refers to the quality of a clinical trial as its 'fitness-for-purpose' (International Council for Harmonisation 2021b). Similarly, the revised Cochrane Risk-of-Bias assessment tool (RoB2) mentions an exception for pragmatic trials regarding certain quality standards. However, the authors barely justify why such an exception is warranted (Higgins et al. 2019). In this chapter, I discuss the concerns regarding the aptness of pragmatic trials to provide unbiased estimates of

treatment effects. In Chapter 3, I continue the discussion of these trials and study the practical value of these designs.

In short, regarding the question of the quality of these designs, I agree with the pragmatists. My discussion shows how the slight shift in purpose between the two types of trials effectively changes the rules of the game. Pragmatic trials can maintain the validity of the causal conclusion, despite a lack of standard controls, by broadening the description of the intervention. However, the manoeuvre comes at the price of informativeness of the causal conclusion, and this situation prompts new epistemic challenges. The first challenge is to meaningfully reconceptualise interventions in pragmatic trials. The second is to establish the practical and ethical relevance of pragmatic interventions for decision-making in healthcare or regulatory contexts. The peculiarities of pragmatic interventions also pose new challenges to the widely held belief that the results of pragmatic trials are easily applicable elsewhere.

To familiarise the reader with pragmatic trials and illustrate what is at stake in the debates, I begin section 1 by introducing two studies – one pragmatic one explanatory – on the effectiveness of the Relvar Ellipta inhaler for treating chronic obstructive pulmonary disease (COPD) as a case study (section 1.1). In section 2, I discuss the epistemic rigour of both types of trials. I compare traditional attitudes to pragmatic attitudes regarding unbiasedness in clinical trials (sections 2.1-2.2). I argue that pragmatists can preserve the validity of causal inferences despite the lack of standard controls by broadening the description of the intervention (section 2.3). This effectively moves the discussion towards the question of how to conceptualise the interventions of pragmatic trials in a sensible way, which I address in section 3. First, I compare the metaphysical ideal underlying effectiveness attributions from traditional trials with that of pragmatic trials (section 3.1); then, I argue that we can adequately conceptualise pragmatic interventions at the level of *therapeutic actions* (section 3.2). Doing so at once reinforces the practical value of pragmatic trials but also uncovers the knowledge loss associated with pragmatic trials. Section 4 is dedicated to the practical use of pragmatic interventions for decision-making in healthcare and regulatory contexts.

1. Pragmatic trials as field experiments

There are many ways to design a randomised trial. The pragmatic–explanatory distinction is an increasingly popular classification of such trials. The distinction dates to a landmark paper in 1967 (reprinted in 2009) by Schwartz and Lellouch,

whose main concern was that most clinical trials were inadequately designed because their design was not aligned with their aim. While most trials aim to inform decision-making (pragmatic), they are in fact designed to increase understanding (explanatory; Schwartz and Lellouch 2009). A similar distinction was made two years earlier by Schneidermann, who distinguished between patient-oriented trials and drug-oriented trials. According to Schneidermann, the former are concerned with the question: 'How shall I treat the next patient [...] who comes into my care?' By contrast, the latter ask: 'Has this drug enough promise that I can bring it into a patient-oriented trial?' (Schneidermann 1966). Since Schwartz and Lellouch's publication in 1967, the pragmatic trial design has been further systematised and multiplied. While the term 'pragmatic trial' is gaining popularity, related or synonymous terms such as 'practical trials', 'large simple trials' or 'naturalistic trials' are also used. I employ the standard terms 'pragmatic trial' and 'explanatory trial' because these are the terms of choice in a well-systematised assessment tool, the Pragmatic Explanatory Continuum Indicator Summary (PRECIS-2) (Loudon et al. 2015). This tool conceptualises pragmatic trials along nine domains, such as recruitment, administration of the treatment and the flexibility of the follow-up. It then ranks them according to how closely the conditions in the trial resemble the practical context in which the experiment is taking place. As its name indicates, PRECIS-2 conceptualises the difference not as a sharp distinction but as a continuum. This is appropriate at the methodological level, because all combinations of rankings in the nine domains are possible practically. In the following discussion, 'pragmatic trial' and 'explanatory trials' mean the extreme ends of the pragmatic–explanatory continuum. In Chapter 3, I introduce the PRECIS-2 tool in more detail for a discussion on the practical relevance of the designs of interest. For the current discussion, the basic principles of a pragmatic trial are sufficient.

I preferably describe pragmatic trials as field experiments of clinical research. Pragmatic trials are embedded into existing infrastructures, processes and resources of the healthcare setting ('the field'). Only those processes and resources that are available to healthcare staff are used, without experimental activities. Hence, pragmatic trials interfere as little as possible in the *natural therapeutic situation*, which means the therapeutic situation as it would take place without the experiment. For example, pragmatic trials are relatively permissive in their eligibility criteria, they do not control how physicians administer treatments or whether patients adhere to these treatments and they do not require the blinding of patients. In contrast, explanatory trials optimally standardise the experimental

situation. The researcher attempts to shield the causal effect of various interferences ('the lab') and therefore artificially alters the natural therapeutic situation for experimental purposes. With pragmatic trials researchers attempt to measure treatment effects 'under routine care conditions' – or natural conditions – rather than under controlled and standardised or somewhat artificial research conditions. To illustrate to consequences of this approach, I compare two trials, one explanatory one pragmatic, as a case study.

1.1. The effectiveness of the Relvar Ellipta inhaler

The Salford Lung Study on the effectiveness of the Relvar Ellipta inhaler is particularly interesting because it illustrates so well the benefits and risks related to pragmatic designs. On the one hand, the pragmatic design of the Salford Lung Study accounted for an important contextual factor in estimating the treatment effect, revealing practically relevant differences between treatments. On the other hand, the pragmatic design in this case favoured the pharmaceutical company GlaxoSmithKline (GSK), which reinforces the critics' concerns that pragmatic studies could undermine the critical standards of evidence. Relvar Ellipta is the brand name of a dry powder inhaler developed by GSK; it contains a combination therapy to treat two different indications, namely, asthma and COPD. Its effectiveness was tested in two pragmatic trials known as the Salford Lung Studies (Vestbo et al. 2016). The trial was discussed as a paradigm case in a stakeholder workshop by the US National Academies of Science, Engineering and Medicine (Downey et al. 2017) because GSK promoted the trials as the first pragmatic phase III pre-marketing study, i.e., conducted before official marketing authorisation of the inhaler was received. Although the trial was a pre-marketing study, the results were not used as evidence submitted for the marketing authorisation. There are also many traditional phase II to III studies available that supported the market authorisation with the European Medicines Agency (EMA; European Medicines Agency 2013). Hence, there is a unique opportunity to directly compare the two approaches in otherwise very similar trials. I only discuss the trial conducted for COPD.

The most interesting aspect of this trial is related to the medicine that was tested, the Relvar Ellipta inhaler. The inhaler contained a combination therapy consisting of the two pharmacological substances vilanterol and fluticasone; the former causes muscle relaxation and the latter decreases inflammation. Vilanterol is a derivate of salmeterol, also developed and marketed by GSK. The major innovation of vilanterol over salmeterol is that it is a *longer*-acting substance, which

reduces treatment administration by half (from twice-daily to once-daily administration). It is a well-established fact that the lack of adherence to treatment is a major obstacle in clinical practice particularly for chronic diseases. Regarding COPD, empirical research indicates that adherence to therapy is below 50% (Lareau and Yawn 2010); research also suggests an inverse relationship between the number of daily doses and the adherence rate (Claxton et al. 2001). The practical advantage of a less frequent administration with the Relvar Ellipta inhaler thus has the potential to increase real health outcomes. To leverage this advantage, the Salford Lung Study did not use any measures to artificially increase the adherence of patients in the study. By contrast, the traditional trial that GSK conducted on the inhaler for submission with the EMA reported adherence rates as high as 97.5% (Agustí et al. 2014).

In addition, an open-label or unblinded design was chosen. In active-controlled treatments with different administration regimens, participants could only be blinded by taking both their active treatment and a dummy placebo treatment that mimicked the treatment of the other group. Such a design would increase the burden of treatment regimens for all participants, and the practical advantage of the inhaler would be neutralised. Thus, the administration regimens differed between the two groups. It was twice-daily for the control group and once-daily for the treatment group, which gave the treatment group a practical advantage in terms of a simplified dose regimen. Table 1 compares relevant aspects of the Salford Lung Study with a traditional counterpart trial that GSK submitted to the EMA for authorisation.[8]

Table 1: Comparing the Salford Lung Study with an explanatory counterpart study

Salford Lung Study (Vestbo et al. 2016)	**Explanatory counterpart study (Agustí et al. 2014)**
Duration: 12 months	Duration: 12 weeks
Intervention: vilanterol/fluticasone combination therapy **Comparator:** usual care (12% mono therapy, 34% dual therapy, 54% triple therapy, for details see table 2 Chapter 3.)	**Intervention:** vilanterol/fluticasone combination therapy **Comparator:** salmeterol/fluticasone combination therapy

[8] It is noteworthy that the explanatory counterpart trial only played a secondary evidential role in the drug licensing process. The primary studies compared the new inhaler to a placebo or to the individual components of the inhaler, while the one I cite here compares the new inhaler to its predecessor. In this regard this traditional study was already slightly pragmatic in the choice of its comparator. The details of the evaluation process appear in the public assessment report by the EMA: European Medicines Agency (2013).

Administration: Treatment group: 1x daily active inhaler Control group: usual care (mostly 2x daily=	**Administration:** Treatment group: 1x daily active inhaler plus 2x daily placebo inhaler Control group: 2x daily active inhaler plus 1x daily placebo inhaler
Primary conclusion: The rate of moderate or severe exacerbations was moderately but significantly lower with fluticasone furoate–vilanterol therapy than with usual care (p = 0.02).	**Primary conclusion:** Improvements in lung function and health status were not significantly different between FF/VI 100/ 25 mg once-daily and FP/SAL 500/50 mg twice-daily.

Perhaps not surprisingly, the Salford Lung Study and its traditional counterpart reached different conclusions. While the Salford study presented evidence of moderate health benefits from the Relvar Ellipta inhaler compared with usual care, the explanatory trial found no such evidence when comparing the inhaler to its predecessor. Interestingly, the investigators in the Salford study were transparent about the fact that they counted on the practical advantage of the simplified dose regimen to make a causal difference:

> [In conventional trials] frequent face-to-face monitoring ensures high adherence to therapy and good inhaler technique. This comparative effectiveness trial that was conducted in a population of patients with COPD was largely unsupervised over the yearlong period, *which allowed important factors in usual clinical care,* such as adherence, frequency of dosing, and persistence of good inhaler technique*, to come into play*. (Vestbo et al. 2016, p. 1260, my emphasis)

In short, the reasoning of the authors suggests that the Relvar Ellipta inhaler is more effective *because it is easier to use*. The Salford Lung Study, with its pragmatic design, translated the practical advantage of the Relvar Ellipta inhaler into the overall effect-size estimate, and this advantage was sufficient to demonstrate superiority over usual care. If we compare the two trials, the standard trial supports the view that there is no relevant difference between the new inhaler and an alternative treatment. The Salford Lung Study, by contrast, implies that there is a relevant difference – which only a pragmatic design could make visible. This difference in perspective lends plausibility to the idea that evidence from a pragmatic trial is practically valuable. If it is indeed the case that patients benefit from the ease of adherence such that it improves their real health outcomes, then clearly decision-makers should consider such evidence.

At the same time, certain methodological features together with the discordant evidence from the traditional trial raise concerns about the quality of

the obtained evidence, for at least two reasons. First, the difference in treatment administration between the two groups introduces a systematic advantage for one group. This point challenges the common understanding of a truly fair comparison, where all contextual factors are kept equal. Second, and perhaps more importantly, while the open-label character of the trial is necessary to leverage the practical advantage, it introduces the risk of placebo effects and other factors associated with unblinded trials. These issues can bias the results.

2. Assessing unbiasedness in clinical trials

It is a common conception that the gain in practical relevance in pragmatic trials comes at the price of epistemic rigour. While there is some truth to this idea, I believe critics have misidentified the problem. Usually, the lack of epistemic rigour is identified as a lack of internal validity or a risk of bias, while I believe that we should understand the problem primarily as an information loss. In the current literature, we can find three views regarding the principled unbiasedness of pragmatic trials. The pragmatists seem to hold that there is nothing wrong, in principle, with the internal validity of pragmatic trials, because they are randomised trials and because they pursue a practical purpose. Zuidgeest et al., for example, argued that 'all other features of such trials are secondary to randomisation and a matter of choice rather than of principle' (Zuidgeest et al. 2017, p. 9). Critics, however, argue that pragmatic trials neglect standard controls, which in any case introduces a risk of bias such as placebo effects, which in turn impairs their internal validity. Fraile Navarro, Tempini, and Teira make such an argument in terms of a lack of 'impartiality' (Fraile Navarro et al. 2021).[9] Taking an ethical perspective, Borgerson argues that regardless of the ongoing debate on the epistemic rigour of pragmatic trials, the gain in social value could potentially outweigh a certain lack of epistemic rigour (Borgerson 2013). If pragmatists want to defend their approach against the well-established traditional epistemology of clinical trials, they need to substantiate their position. They need to show, first, why experimental controls that are well-established standards are 'a matter of choice' in pragmatic trials. Second,

[9] Their argument has a historical and social perspective. Within that perspective, 'impartiality' of a method is relevant to promote the acceptability of results among experimentalists. Yet, on the methodological level, I understand that impartiality implies the same operationalisation and epistemic rationale for experimental controls as the standard epistemology does for unbiasedness.

they need to show what they can gain from making such choices in epistemic and practical terms.

I develop a line of argument that explains how pragmatists attempt to preserve the internal validity of their causal inferences through broadening the scope of the intervention that is being tested. This echoes with some assertions found in the literature. Schwartz and Lellouch asserted that pragmatic trials have somehow broader notions of the intervention that 'absorb' the context in which they are administered (Schwartz and Lellouch 2009). Bluhm recognises that pragmatic trials are not as good at 'isolating' the treatment (Bluhm 2017, p. 100). However, they do not provide a systematic explanation what these expressions mean and what the epistemic and practical consequences of such treatment definitions are. My contribution closes this gap.

2.1. The traditional EBM attitude towards unbiasedness

Today's standard epistemology of clinical trials is a result of the EBM movement, which advocates for the noble goal of supporting all healthcare decisions with the best available evidence. The main activity of these communities is to publish systematic reviews that amalgamate the best available evidence in the service of clinicians. Their biggest success has been the wide uptake of their hierarchical theory about the quality of evidence, the well-known evidence hierarchies that place RCTs at the apex. The most influential organisation in the EBM movement, the Cochrane Collaboration, additionally requires assessing individual randomised trials with a quality assessment tool that examines the individual *risk of bias*, to weight the trial's impact on the overall evidence synthesis. I agree with Stegenga that this is a worthwhile undertaking, because method types are insufficient to support the quality of individual instantiations (Stegenga 2018, Chapter 5; for a discussion of quality assessment tools, see Stegenga 2018, Chapter 7). Hence, in these quality assessments, pragmatic trials can run into trouble regarding judgements about their quality of evidence.

Quality assessment tools look at the individual features and the detailed execution of a trial, such as the randomisation process, allocation concealment, blinding, adherence control and drop-out rates. These tools come in various forms, with different levels of complexity. One of the most detailed developed tools, the Cochrane Risk-of-bias (RoB) tool developed a rule-based decision framework (Higgins et al. 2019). The tool ranks the individual experiments according to their *risk of bias*, where bias is considered a distortion of the estimate relative to the true

treatment effect. Depending on the tool that is used, it could alert us to several risks of bias in the Salford Lung Study, such as the following:

1. Participants were not blinded as to which treatment they received.
2. Adherence of participants was not controlled.
3. Patients in the treatment group were allowed to change to the control group at any time.

A critic who questions the quality of the Salford study could point out that the study lacked standard controls and therefore incurred a risk of bias, which is a sign that clinicians should not rely on the results produced by such trials. Fraile Navarro and colleagues, for example, criticise the Salford study precisely on the ground that it was unblinded (Fraile Navarro et al. 2021).

Philosophically, the role of experimental controls that are captured in these quality assessment tools is intuitively explained by using Mill's method of difference and the principle of eliminating alternative hypotheses. Such controls are crucial in the trial because they make the treatment and control groups comparable regarding relevant factors. This equality allows for eliminating those factors from the list of possible alternative explanations. These factors include the placebo effect, which could have caused clinical improvements; the experimenter, who could have intentionally selected the patients; or patients dropping out from a study, among others.

In Chapter 1, I distinguished between precision and validity and argued that measures to increase homogeneity in a Millean sense are meant to increase the experiment's precision and not its validity. The distinction between validity and precision is highly relevant in this context because many experimental controls that pragmatic researchers drop are controls that pay into the precision of the estimate but not its validity. For example, pragmatic trials include patients of diverse ages, diverse administration practices, diverse practical settings with diverse disease stages or diagnoses based on different criteria. The lack of such standardisation measures must be compensated by an increased sample size to reach sufficient precision, but it does not diminish the experiment's validity. Quality assessment tools like the Cochrane RoB-2 build on this distinction and generally do not include checks for experimental measures that are meant to increase precision.[10]

[10] As part of risk-of-bias assessments, researchers sometimes check for 'imbalances' of causal factors in the characteristics of patients. We should understand this task as aiming to detect potential signs of failed allocation concealment in the randomisation process. This is, of course, an imperfect indicator for a failure of allocation concealment, because it cannot distinguish between an imbalanced but random allocation and an intentionally imbalanced allocation. To recall the relevant epistemic difference from the frequentist's

These tools are employed to make risks visible that would otherwise remain invisible in the causal inference. Following Baetu, all practices that increase the precision in an experiment can be understood as eliminating the hypothesis that an effect occurs by chance alone (Baetu 2020); however, the uncertainty implied by the chance hypothesis is already made visible in the uncertainty estimate produced by statistical significance tests. Hence, the absence of controls in pragmatic trials that increase precision can be neglected because they are not required for unbiased causal inference and already made visible in the value of statistical significance.

Not all controls that pragmatic trials abandon are like that. Among the others, the lack of blinding is arguably the greatest threat for pragmatic trials. The unavoidable consequence of omitting such controls is simply that the experimental setup does not provide an immediate justification to rule out certain alternative hypotheses. If a trial lacks blinding, for example, we cannot rule out the hypothesis that the placebo effect has caused the benefit, or that patients in the control group sought additional treatments. These risks weaken the support for the cause–effect hypothesis of interest, namely that the treatment and not something else caused the benefit.

The standard strategy to deal with such a risk is to use background knowledge and evaluate the chance of a certain bias occurring in a particular case. For example, Howick argues that large effect sizes of highly effective medicines count as a reason that blinding can be unnecessary (Howick 2011, chapter 6). Similarly, the methodological literature often holds that objective outcomes, such as mortality, are relatively unlikely to be affected by a lack of blinding. Within the standard approach to unbiasedness, such additional assumptions can count as reasons to (cautiously) neglect the influence of certain biases. Such assumptions are arguably more difficult to defend than the successful implementation of blinding because they rely on rare events and reliable background knowledge. Therefore, stakeholders usually only tolerate such indirect approaches in cases where blinding is either impossible or unethical. Hence, the critical argument regarding the lack of unbiasedness in pragmatic trials remains: Given that blinding (and other controls) are generally acknowledged as standard quality measures to ensure unbiasedness, the omission of blinding implies a prima facie risk of drawing false causal

perspective: the random allocation has a known probability distribution, while the probability distribution of a purposeful allocation is anyone's guess. From a Bayesian perspective, it makes no relevant epistemic difference how the allocation came about.

conclusions. To the extent that pragmatic trials are unblinded and avoid other controls, we run a higher risk of drawing false causal conclusions.

2.5.Pragmatic attitudes towards unbiasedness

How can pragmatists reply to the threat that the omission of blinding and other controls poses to their work? First, they can employ the kind of reasoning according to background knowledge mentioned above. For example, it is not unusual for pragmatic trials to employ objective outcomes, such as mortality or the use of healthcare resources, to minimise biases in the subjective evaluation of outcomes. In the Salford Lung Study on COPD, the primary outcome was acute exacerbation of COPD, which is objective in the sense that these outcomes are not reports of how patients evaluate their own wellbeing; they are based on physical symptoms that are externally evident. These outcomes are also robust in the sense that symptoms are so severe that they require patients to contact the treating physicians. The risk of events going unnoticed because patients decide to care for their symptoms by themselves is relatively small. Hence, Fraile Navarro and colleagues might be right when they criticise the Salford Lung Study on asthma was unreliable because the primary outcome was based on subjective outcomes from the asthma control test (Fraile Navarro et al. 2021), which includes items such as 'Asthma keeps you from getting much done at work/school' (Nathan et al. 2004, p. 62).

Second, the pragmatist can adopt a different attitude: The pragmatist can argue that they do not need to eliminate certain alternative hypotheses at all, because this serves their research interest. I cite again the case of blinding to assess the success of such an argument. Blinding is used to put participants in a state of ignorance with a twofold epistemic role. If implemented successfully and together with a placebo control, a trial can estimate only the effect of the 'characteristic features' of an intervention, i.e., those that are assumed to play a causal role in the mechanism of action (Howick 2017). Blinding also ensures proper randomisation by preventing protocol violations resulting from awareness about treatment allocation, such as patients dropping out from the trial or seeking other medication outside the trial (Howick 2011; Teira 2013, for a discussion see Chapter 1).

Regarding the first role of blinding, namely eliminating the effects of attitudes towards a treatment, the pragmatist's case is not complicated. They acknowledge that patients' expectations can modify what the patient believes about a treatment and thus affect how the patient benefits from a certain treatment. What they want to learn from the experiment is how patients benefit from healthcare decisions

'under the conditions of routine care', and these conditions include the effects of patients' attitudes about the treatment. Consequently, pragmatic trials estimate the effectiveness of treatments, including placebo effects. Pragmatists can, if they want, embrace the claim that this is precisely what they attempt to measure:

> In pragmatic trials, as in the real world delivery of care, blinding of participants and clinicians may be impossible. Belief (or disbelief) in the intervention, extra enthusiasm and effort (or less), and optimism (or pessimism) in the self-assessment of outcomes may thus add to (or detract from) the effects of an intervention. Pragmatic trials may incorporate these factors into the estimate of effectiveness, rendering the findings more applicable to usual care settings. (Zwarenstein et al. 2008, p. 6)

This strategy is conceptually consistent. If pragmatists embrace this line of reasoning, placebo effects are no longer a bias in the study but become part of the effect that is being measured. If they stand by their word – that such estimates of effectiveness render 'findings more applicable to usual care settings' – they could even argue that blinding in pragmatic trials is not only impossible but also undesirable.

To explain why pragmatic trials might also not require blinding to prevent protocol violations requires a bit more work. In Chapter 1, I argued that this problem affects trials differently depending on how well they are aligned with the preferences of patients. If assignment to the control group contradicts their interests, they are more likely to act upon their interests and break the protocol. The interests of patients depend on the severity of their disease, its progression, alternative treatment options and the effectiveness of the active treatment. The more treatment options appear equivalent for individual patients, the less severe is the risk that they might break the protocol to act upon their own interests. In a trial of pragmatic nature, patients usually receive the standard of care, which has a good chance to be aligned with their interests. Hence, unlike in some placebo-controlled trials, the basic cooperation of patients is highly plausible.

As I argue in Chapter 1, all clinical trials can tolerate some protocol violations, particularly those taking an 'intention-to-treat' approach for the analysis. Pragmatic trials are particularly tolerant for these types of behaviours. Of course, patients in pragmatic trials still have certain preferences, expectations or habits regarding specific treatments. These can prompt certain behaviours, such as patients dropping out from the trial because they preferred their old therapy or seeking additional treatments because they are unsatisfied with the outcome of the trial. However, the same types of behaviour also occur under normal treatment conditions. To the extent that these are *natural* behaviours of patients, the same

reasoning applies to these factors as reasoning about the problem of placebo effects. These behaviours do not need to be controlled if we aim to estimate the treatment effect under routine conditions. Indeed, we certainly would not want to control them. In short, because pragmatic trials have a different epistemic aim, these types of behaviours are formally allowed by the protocol. They just represent the type of behaviours that patients typically do – which sometimes means that they stop treatments, switch treatments or seek additional care from elsewhere. Hence, pragmatists are right in claiming that changing the purpose of the trial also changes the rules of the game.

The pragmatic attitude towards blinding applies to other experimental controls, too. For example, the Salford Lung Study wanted to reflect the adherence of patients as it would happen without the experiment. Therefore, the systematic difference in the treatment regimen did not introduce a bias into the experiment but mainly reflected the research interest. Similarly, the risk that patients misdiagnosed with COPD might have been included in the Salford study, is not a genuine risk if it reflects the fact that those patients receive the treatment in routine care. In other words, if we are interested in the benefit that patients gain from therapies in routine care conditions, controlling for contextual factors – even if they are systematically different between the compared groups – does not necessarily increase the validity of the experiment. Quite the opposite can be the case.

The pragmatic attitude towards biases is, however, a nuanced matter. Pragmatists can run into problems with biases if the trial introduces systematic differences in beliefs and behaviours that solely occur because of the experiment. Examples are 'being in the control group' or 'receiving an experimental medicine'. In other disciplines, such effects are known as the Hawthorne effect. They occur if research participants form expectations about the experimental procedure and react to those expectations (Teira and Reiss 2013). These effects could lead patients to report outcomes that are not actually occurring because they are in an experiment and not because their reports reflect their natural behaviour. Such effects are a clear threat to the pragmatic attitude towards biasedness; pragmatists do need to minimise this threat to achieve their goal.

A choice for objective outcomes might be crucial in this regard. Interestingly, pragmatism itself also has a role to play. The pragmatist's efforts to seamlessly embed the experiment into the routine care processes is an attempt to make patients *indifferent* to their participation in a study and eliminate any strong expectations about the experimental procedures. To support that goal, certain trial designs and consent procedures have been developed to withhold information

from patients about their participation in a control group.[11] In a similar vein, there are ongoing discussions about whether pragmatic trials with minimal risks could neglect the requirement to seek informed consent (Kalkman et al. 2017b; Faden et al. 2014). To the degree that pragmatic trials can accomplish their mission to make patients indifferent, pragmatists have grounds to neglect such risks. Certainly, this requires a careful conduct of the trial, similar to successful blinding in traditional trials. Furthermore, some argue that the use of real-world data comes with 'a specific beauty' that outcome collection by physicians is formally blinded because the physicians who collect the data are generally independent from the research endeavour (Hemkens 2018). Other scholars have pointed out that it might still be possible and desirable to blind assessors in pragmatic trials (Zwarenstein et al. 2008).

My point is not that pragmatic trials never carry a risk of bias, nor that we should embrace the pragmatic approach in all circumstances. The point is simply to illustrate that the considerations that are relevant to assess the risk of bias in pragmatic trials differ from those that are relevant in assessing such risks in traditional trials. Hence a shift towards a pragmatic purpose of the trial changes the rules for quality assessments.

2.2. The costs of unbiasedness in pragmatic trials

A common statement about the epistemic characteristics of pragmatic trials is along the lines of 'pragmatic trials increase external validity at the expense of internal validity'. In the previous section, I argued that pragmatic trials have different rules to assess their internal validity or unbiasedness. However, this comes at the price of invoking a broader notion of the treatment. Turning back to the historical origins, we find that Schwartz and Lellouch drew attention to this fact early on:

> The basic principle that two treatments must be compared in two groups which are in every other respect comparable is in no way contradicted by optimization of the contextual factors. Instead, *these factors become themselves part of the therapies* to be compared and are thus distinguished

[11] Such trials, also called registry-based trials or TwiCs (Trials within a Cohort), sample a subgroup of patients within an observational cohort and randomly allocate the selected patients to a treatment group and a control group. However, only patients who are allocated to the new treatment must be asked for consent to receive the new treatment. For patients in the control group, nothing changes. Because all patients consented to these procedures upon their entry in the cohort, the control group does not need to be re-consented, hence they do not know about their participation in the control group until the end of the experiment (when they are usually informed) James et al. (2015). Finally, for retrospective observational studies using real-world data, the concerns about such artificial experimental effects completely disappear.

> from non-contextual factors for which comparability must be assumed. It is characteristic of the pragmatic approach that the *treatments are flexibly defined and 'absorb' into themselves the contexts in which they are administered*. (Schwartz and Lellouch 2009, p. 500, my emphasis)

By 'optimisation' of the contextual factors, the authors mean that some factors can be tailored to the control and treatment groups separately rather than being equalised between them. In the Salford Lung Study, for example, administration of the treatments was tailored to the groups separately by allowing the treatment group to take the medicine once daily and the control group to take it twice daily. The following two phrases in the quote, although slightly obscure, point to the conceptual consequence of such a design choice: 'these factors become themselves part of the therapies'; and the idea of treatments which 'absorb into themselves the contexts in which they are administered'. A causal explanation draws a distinction between *the* cause (the medicine) and its causally relevant background conditions (e.g. adherence to the medicine). To clarify their wording, I suggest understanding this as a shift in the distinction between the primary cause and its background conditions. In a pragmatic trial, all causes that are not equalised between treatment and control group are not part of the background conditions but instead included in the conceptualisation of the primary cause, that is the medical intervention. This is consistent with the observation from the Salford Lung Study. What is usually deemed to be a mere background condition – the adherence of patients to a therapy – was allowed to play an active causal role in the study. In other words, since these experiments cannot and do not attempt to equalise certain background conditions, the conditions can be reinterpreted as primary causes contained within the intervention.

This conceptual shift is key to understanding the pragmatist's attitude towards unbiasedness. Pragmatic trials can fulfil the following plausible explication for internal validity – which I believe is in line with one relevant use in the methodological literature:

IV: An internally valid experiment supports a causal inference from the observed effect to the intervention as its cause.

In the Annex, I further elaborate and defend this definition. Pragmatic trials can support an inference from the observed effect to the intervention as its cause, because the trial includes all unequal factors within its notion of the intervention. Likewise, controls other than randomisation are a matter of choice if pragmatists are willing to pay the price of a broader notion of their treatment. In other words,

all other features than proper randomisation are non-essential properties for valid causal-inference in randomised trials.

These findings also illustrate what the true costs of pragmatic trials are. Valid causal inference in these trials comes with the price of an information loss about the contribution of any of the individual causal factors contained within the broad notion of the intervention. Particularly, they come at the price of an information loss about the causal contribution of the pharmacological properties of the drug. Hence, pragmatic trials cannot fulfil the requirements of another widespread use of the notion of internal validity, namely where internal validity is preserved for causal inferences that allow an inference to a single causal variable.[12] Pragmatists would be fighting a losing battle if someone were to insist that this narrow sense of internal validity is the only one. However, this battle would not even be worth fighting, because the property of internal validity then just becomes irrelevant to pragmatists. What is at stake in the debate is the risk of drawing false causal conclusions – and not the number of causal variables involved.

I hold that the perceived lack of epistemic rigour in pragmatic trials is better understood as an information loss about individual causal variables, rather than as a risk of drawing false causal conclusion about the intervention, broadly conceived. Hence choosing between pragmatic and explanatory trials is not a matter of different risks for drawing false positive or false negative causal conclusions, which philosophers call inductive risks. Bluhm used the example of pragmatic clinical trials to argue that the inductive risk perspective is helpful but insufficient to describe the epistemology of different types of clinical trials, and I agree (Bluhm 2017). Rather it is, as Schwartz and Lellouch made clear early on, a question of choosing the right design for the right purpose.

My comment is intended as a principled point to lend plausibility to the pragmatist's unconventional view. This principled point can motivate the position that the difference in purpose between pragmatic and traditional trials necessitates different quality assessment rules – where blinding and other controls do not play an identical role as in traditional trials. Furthermore, pragmatism itself can become relevant to ensure unbiasedness. However, it is the burden of the pragmatists to clearly define such rules and develop a suitable and rigorous quality assessment

[12] Campbell's dissatisfaction with this particular usage of the term led him to rename the terms in his later work, as this usage was not what he envisioned when he coined the term in the 1950s (Campbell (1986)). His new term of art was 'local molar validity', which evaluates the question: 'Did this complex treatment package make a real difference in this unique application at this particular place and time?'

tool to address risks of biases that are unique to pragmatic trials. The current state of the methodological literature does not fulfil this requirement. For example, the Cochrane Risk-of-Bias 2 tool mentions an exception for pragmatic trials regarding the need for blinding because trials measure 'intervention strategies of individuals who are aware of their care' (Higgins et al. 2019, p. 23). Yet the authors barely justify such an exception or explore other risks that can occur in pragmatic unblinded trials (as discussed above). The state-of-the-art tool for pragmatic trials, the PRECIS-2 tool (Loudon et al. 2015), allows researchers to assess the degree of pragmatism of a trial – yet it does not address the relation between pragmatism and unbiasedness. Neither does it address the impact of blinding on either of these dimensions. In fact, pragmatists do not claim that pragmatic trials are necessarily unblinded; it just seems to be implied by their attitudes. I have demonstrated the direction in which such an assessment could go. The other project that pragmatists need to undertake is to provide a theory of medical intervention that pragmatic trials measure and demonstrate that such notions of the intervention truly are practically more useful than interventions measured in conventional explanatory trials. I illustrate what such a theory of pragmatic interventions could like in the next section.

3. Conceptualising pragmatic interventions

3.1. Hunting true treatment effects

Theories about hypotheses that are tested in clinical trials have mostly focused on distinguishing between different levels of outcome measures (Stegenga 2018, chapter 8) or the range of background conditions within which the obtained results hold true (Cartwright 2012). It seems there is little of philosophical interest in the fact that experiments can test different interventions. The previous discussion however has shown that the medical interventions tested in pragmatic trials differ markedly from the subject matter of traditional trials. The former interventions contain causal factors that conventional trials conceptualise as risky biases or background factors. The shift between the primary causal factor and its background conditions, I believe, creates a tension with what we mean when we attribute effectiveness to medical interventions. This tension moves the discussion towards the question of whether we can conceptualise the interventions of pragmatic trials in a sensible way.

According to Stegenga's hybrid theory of disease, medicines are said to be effective if they either target the causal basis of disease – the biological dysfunction – or the normative target of disease – the harms. Moreover, medicines can target different levels of disease. They can, for example, target the physiological mechanism or the clinical symptoms (Stegenga 2018, chapter 3). Ashcroft's analysis of clinical effectiveness focuses on what it means to attribute effectiveness to medical *interventions* and he attempts to get to the metaphysical grounds of the notion (Ashcroft 2002). He argues that clinical effectiveness is a therapy's 'capacity to φ', understood as a property of a therapy. In his view, such a capacity can be further analysed as a therapy's function to *φ*, which supervenes on its 'intrinsic' physical features. For example, when we say that aspirin has clinical effectiveness, we mean that aspirin has the capacity to relieve headaches by virtue of its intrinsic physical properties (which are to inhibit the enzymes COX-1 and COX-2). Ashcroft's analysis fits well with what Schwartz and Lellouch identified as the *explanatory* attitude in traditional trials, i.e., an attempt to establish a narrow physiological hypothesis and isolate the effect of the medicine.

While the intrinsic properties of a medicine that are directly involved in the mechanism of action are of special interest in clinical trials, medicines have various other properties that can determine how effective they are. For example, medicines come with side-effects, a treatment schedule and a route of administration – and with a certain taste, colour and shape. They even involve, in the wider social context, an image, a supply chain, a reimbursement plan and a healthcare delivery system. It is the effect of this rich causal nexus of medicines and the various interactions between healthcare agents and social context that is the research interest of pragmatic trials. As Ashcroft's analysis implies the special interest in the pharmacological properties is generally justified by the idea that the pharmacological properties involved in the mechanism of action are *intrinsic*, whereas others are merely accidental. I do not think that such a clear-cut distinction can be drawn. In the case of the Relvar Ellipta inhaler, the practical advantage of the medicine was not a mere contingency of the medicine. The medicine's simplified administration supervened (in Ashcroft's words) on the medicine's intrinsic physical properties of being a longer-acting substance. The same holds for most of a medicine's side-effects, which cause different interactions with the medicine. Even properties that appear purely contingent somehow supervene on intrinsic properties. The supply chain of a medicine, for example, supervenes on the physical properties that determine when a medicine expires or at what temperature it needs to be stored. Sunscreen that would not expire after a year

could certainly prevent more sunburns than a conventional sunscreen. Importantly, these properties are not only additive to the genuine effect of the medicine but can modify the treatment effect in highly interactive ways. For example, inhalers that prompt a quick response could motivate patients to adhere to a frequent administration, which would increase the long-term effectiveness.

I think it is an idle question which of these effects is more genuine and real or even the *true* treatment effect. This point applies to the critics as well as the advocates of pragmatic trials. Critics widely share the intuition that there is something particularly genuine about the effects of well-controlled trials. By contrast, advocates of pragmatic trials refute this intuition with the rhetoric of generating evidence about the 'real world' or 'real patients'. Indeed, the entire real-world evidence movement exploits such rhetorical persuasion. In both cases, patients either genuinely benefit from a certain treatment or do not. In one case they benefit from the isolated effect of the medicine, while in the other case they benefit from the treatment with various interaction effects. I propose that the considerations above establish that both effects are legitimate subjects of clinical investigation.

Indeed, because of their flexibility on the level of the intervention, pragmatic trials have the advantage of accommodating more complex interventions more easily. For example, a common argument states that complementary and alternative medicine (CAM) treatments cannot adhere to the standards of EBM because they rely on highly individualised and holistic notions of medical interventions. That is, CAM treatments are often described as including the patient's relationship to practitioners as well as self-healing effects such as the placebo effect, and the interventions are not identical for any two patients. In other words, advocates of CAM treatments argue that the distinction between intrinsic and accidental features of a treatment is not meaningful in the context of CAM treatments. They reject a method that relies on such a distinction as inadequate to evaluate the effectiveness of CAM treatments; hence they reject the meaningfulness of the randomised trial (Ernst 2002; Borgerson 2005; Tonelli and Callahan 2001). Defenders of the EBM paradigm have countered, and I think rightly so, that pragmatic clinical trials are indeed well suited to accommodate such unconventional therapeutic approaches (Hansen and Kappel 2010). For example, a pragmatic clinical trial could answer the broad question: 'Can assignment to a CAM practitioner increase patients' wellbeing?' Or 'Can assignment to a CAM practitioner reduce patients' use of traditional medical resources?' The intervention here is conceptualised in the broadest possible terms. From such a

trial scientists cannot infer that something *intrinsic* to the CAM treatment reduces the use of other medical resources. Nonetheless, the effect of such an intervention is not less real because we cannot attribute it to a precise individual causal variable. Yet, the challenge remains to provide a theory of medical interventions that would allow for agreement about what a pragmatic trial measures and to develop the necessary conditions for valid causal inference. Below I propose what such a theory could look like.

3.2.Interventions as therapeutic actions

When Schwartz and Lellouch introduced the pragmatic–explanatory trial distinction into the literature, they proposed distinguishing between the effects of drugs and 'treatment strategies'. The former is the subject matter of traditional trials; the latter is of interest in pragmatic trials. The distinction indicates that a treatment strategy goes beyond the mere choice of a single therapy, suggesting that in healthcare, one often opts for a combination of therapies. Today, the main tool to explain the difference between pragmatic and explanatory hypothesis is the efficacy–effectiveness distinction (Eichler et al. 2011; Nordon et al. 2016). According to this characterisation, the difference lies in the background conditions under which the effect of a therapy holds. The efficacy–effectiveness distinction, although commonly employed, is an inadequate tool to conceptualise the difference between the two types of trials. The distinction implies that the intrinsic properties are the most important causal drivers; everything else is subsumed under the notion of (unimportant) background conditions that do not need to be specified further. The distinction does not capture that the intervention itself has a broader scope. An adequate theory should at least partially specify what the causal factors are that are 'absorbed' into the intervention; more importantly, it should set boundaries as to what counts as a successful pragmatic intervention.

My proposal to refine pragmatic interventions is to conceptualise them at the level of *therapeutic actions*. By 'therapeutic action' I mean, for example, 'to prescribe therapy X', 'to administer therapy X', 'to take medicine X' or 'to recommend a health behaviour Y' and so on. This concept of medical interventions allows complex causal interactions to take place at a low level of description and reflects appropriately what we can and cannot learn from pragmatic trials. Most importantly, the suggestions reflect that, in many cases, researchers remain ignorant about the lower-level causal description and the precise contribution of such causes, particularly the contribution of characteristic features

of the drug. At the same time, the description can be adjusted to reflect different degrees of experimental controls. For example, the action 'taking medicine A' implies that adherence can be assumed (the medicine has been taken). It does not rule out that the act of 'taking' has played a causal role in the effect (it was not blinded and thus did not eliminate placebo effects). Likewise, the action of 'prescribing medicine A' does not determine how the causal story continues after the action of prescribing the therapy has occurred; the patient could go home and throw the prescription in the garbage. This intervention captures what we can learn from a trial with no control or follow-up of adherence. Reflecting the level of control in the conceptualisation of the intervention prevents our being led astray in our causal inferences, particularly in terms of attributing the causal efficacy to the treatment alone (which clearly risks being an erroneous inference).

Bringing this logic together with the previous section, it follows that pragmatic trials support valid causal inferences; however, these inferences relate to therapeutic actions rather than the pharmacological properties of a medicine. When adequately designed and conceptualised, pragmatic trials can also underpin meaningful causal conclusions. Note this proposal does not suggest that the medicine can be eliminated or replaced altogether in the therapeutic action. The comparative and randomised nature of the experiment supports that it is the prescription of *this* medicine, rather than another, that has caused the observed effect. What is at stake is (only) which of the properties of the intervention precisely contributed to the effect. In an unblinded trial, the difference might be caused by patients' expectations about *this* medicine rather than the pharmacological properties of *this* medicine. Nonetheless, some property of the medicine – even if only the property of being new on the marked – must be involved in the effect.

In addition, a theory of pragmatic interventions must set limits to meaningful and practically useful pragmatic interventions. A pragmatic intervention of 'prescribing medicine A' might cause side-effects motivating patients to seek additional treatment, which is in fact the direct causal driver of the health benefit. In this case, it seems undesirable for our practical interests to conclude that 'prescribing medicine A' is a good therapeutic decision. To avoid such undesirable conclusions, pragmatists should set boundaries to the tolerated natural behaviour and measure these behaviours in their trials as an indirect form of control. If researchers find that the behaviour of patients exceeds these thresholds, they should conclude that the implementation of the medical action failed. To that end, a better theory of therapeutic interventions in pragmatic trials is indispensable;

otherwise, we risk only multiplying the evidence base without agreeing on what the evidence is about.

4. What is the use of pragmatic interventions?

In the first three sections, I have substantiated the pragmatists' position by developing the conceptual scaffold to better understand causal inferences in pragmatic clinical trials. I have shown that pragmatic trials can maintain the validity of the conclusion simply through broadening the scope of the intervention. I proposed that we could try to conceptualise such interventions in terms of medical actions. Understandably, critics are hardly convinced by the pragmatist's move because it does not respond to the driving factor behind their concerns. That is, it does not respond to the concern that pragmatic evidence could increase the risk that treatments are prescribed or authorised mainly because of placebo effects or other contingent contextual factors. Rephrasing the problem in terms of an information loss rather than validity does not eliminate this risk. A second concern is that the information loss in pragmatic trials is practically relevant for treatment decisions because making sensible treatment decisions requires knowledge about the effects of stable causal factors and not accidental contextual factors. Chapter 3 is dedicated to the question of extrapolation in pragmatic clinical trials, so I postpone discussing the second concern to the next chapter.

I see three potential responses to the concern about placebo effects. The bold pragmatic response to the critic's concern could reiterate the pragmatic attitude from section 2.2: As long as the health benefits are real, it does not matter what exactly caused such benefits. What matters is *that* we help patients, not *why*. Coming back to the example of CAM treatments, one could insist that it really does not matter whether patients are feeling better because they have a healing relationship with the practitioner, because they believe in and activate their self-healing capacities or because needles were inserted into the body. What matters is that patients are doing better as a consequence of their visits to the CAM practitioner. A similar claim could be made about our case study. The Relvar Ellipta inhaler seems to be moderately more beneficial than patients' usual care. There are two potential explanations for this observation. One is that the benefit was caused by the ease of administration, which increases natural adherence to the medicine and hence its effectiveness. The other explanation is that the moderate benefit is caused by the placebo effect or other behaviour that resulted from

patients' awareness about the treatment.[13] The pragmatist might hold that the difference does not really matter; in either case, the Relvar Ellipta inhaler improves health outcomes and seems the best treatment option. One might even argue that it is only due to the pragmatic perspective that we can make the best treatment decisions, which is to prefer the most effective therapeutic action.

While there is some appeal to this argument, it is a bold move to defend such a position in all circumstances. The position seems to imply that we are ready to accept that we might expose patients to *unnecessary* harms. The following assumptions give rise to this concern. First, it is plausible to assume that relevant harms are primarily caused by the characteristic features of a therapy rather than contextual factors. Second, if the benefit is actually not caused by the characteristic features, we could imagine swapping the harmful properties of the medicine with something else – a placebo – to preserve the benefit, while eliminating the harms. Hence, exposing patients to these harms would be unnecessary and therefore hardly a good treatment choice. If we want to make ethically justifiable treatment decisions, we need to know that the harms of a treatment *necessarily* accompany its benefits – that is, both aspects are causal effects of the characteristic features of the medicine. Given the difficulties in distinguishing intrinsic versus contingent features of a medicine (see section 3.1), one might doubt whether these assumptions truly hold. However, I accept that the ethical weight of these concerns outweighs the problem that the underlying assumptions are idealised.

The usual response by pragmatists to these conncerns is to limit the use of pragmatic trials to situations where the risk of unnecessary harms can be ruled out. In current practice, this is ensured by assigning pragmatic trials the role of post-marketing studies. They are conducted after an explanatory study has demonstrated effectiveness under controlled conditions for regulatory decision-making; complementary to explanatory trials rather than being in competition. Moreover, pragmatic trials are still clearly preferred to evaluate non-pharmaceutical interventions instead of drugs (Hirt et al. 2024). However, with the emergence of the real-world evidence paradigm, this situation is changing, and pragmatic trials and real-world evidence are increasingly of interest also for regulatory approvals.

[13] In any randomised trial there is always a third alternative explanation, namely that the trial has not been well-conducted and hence the measured effect is unreal. Following the medical nihilist's position, this might even be the most likely explanation. In my view, whatever the likelihood of this third explanation is, such likelihood is comparable in pragmatic and explanatory trials.

In the regulatory context, without prior evidence that allows regulators to rule out the risk of unnecessary harms, it seems there are two options. The first option is to question the relevance of placebo effects in the bigger picture. An important motivation behind the pragmatic stance is the discrepancy between the health benefits found in clinical trials and the benefits that actually occur in healthcare. What draws the attention of practitioners to this problem is not that moderate effects in clinical trials demonstrate greater therapeutic power in practice. On the contrary, in most cases, the discrepancy is expected to run in the opposite direction. Hence, so the assumption goes, in most cases pragmatism can counteract an interest in bringing to market a therapy that shows only a moderate effect. In turn, this would eliminate from clinical practice those substances that cannot bridge the efficacy-effectiveness gap. In this bigger picture, the threat that a few exceptional cases nevertheless may introduce the risk of unnecessary harms then is a bullet one could be willing to bite. If this intuition is true (or we could at least identify instances where it is true), this rationale could undergird the pragmatic approach with considerable epistemic value. However, this notion is mostly based on intuitions rather than reliable empirical evidence. It is a widespread assumption that pragmatism yields, overall, smaller effect sizes than placebo-controlled explanatory trials. If this is true, we could assume that the factors that decrease the effect sizes of treatments in natural conditions outweigh the factors that increase it, like placebo effects. However, meta-research on how pragmatism influences effect estimates is still scarce, and the little that exists is rather inconclusive. I come back to this proposal in Chapter 3.

The second option is to counterbalance the risk by raising the standards for a sufficient benefit–risk balance. Stegenga has convincingly shown that small effect sizes are common in clinical trials and create the problem that we cannot distinguish between biases and treatment effects (Stegenga 2018, chapter 11). The same holds for distinguishing between placebo effects and other causes of the treatment effect. A simple way to adjust this balance is by increasing the threshold that defines the minimal clinically important difference. This measure defines the minimal change in a treatment outcome that is relevant from a clinical perspective. A successful clinical trial needs to meet two thresholds, i.e., the minimal clinically important difference as well as statistical significance. By raising the former threshold, placebo effects become a less convincing explanation for the treatment effect. Such a solution would shift our focus from measuring (theoretically hard to justify) true effects toward clinically relevant effects that matter *despite* placebo effects. This solution has the advantage of being equally applicable to both

pragmatic trials and conventional trials that carry a risk of unsuccessful blinding. However, it entails the difficulty that there is no obvious answer to how large such an effect size must be, particularly in active-controlled trials, where the effect estimate depends not only on the effectiveness of the treatment but also on the effectiveness of the control. Following Stegenga, the industry mostly develops 'me-too' drugs. These are treatments that belong to the same class of medicines as already available treatments; the Relvar Ellipta inhaler is a good example of such a drug. It would be surprising if 'me-too' drugs were markedly more effective than other members of the same class (Stegenga 2018, chapter 4). Moreover, this solution is in tension with the first one above, stating that effect sizes in pragmatic trials further diminish, and it is not evident how this tension can be resolved.

Let's take stock. The choice between pragmatic or explanatory attitudes implies a shift in perspective. The traditional epistemic question regarding which of these two methods produces more true results, however, does not make sense. Because these trials measure the causal effects of different therapeutic entities, we cannot compare the false-positive and false-negative rates between the two types of trials in an attempt to fit them into the EBM evidence hierarchies. Instead, the choice between these trials is determined by the goals that need to be achieved. The good news is that clinical trials can deal with a wider range of problems than the randomised method is currently credited with doing. That is, randomised trials do not rely on homogeneous populations, placebo controls or blinding to provide unbiased causal knowledge. These properties are, from the pragmatic perspective, non-essential. They also can deal with unconventional interventions such as CAM treatments. An interesting upshot of my arguments is this: The epistemic and practical merits of pragmatic trials demonstrate that many of the widely criticised limitations of randomised trials are in fact limitations of explanatory trials. In other words, arguments in favour of observational studies are often based on a false dichotomy. They imply that our only choice is between explanatory and randomised trials (blinded, placebo-controlled, highly selective, narrow interventions) versus observational studies (which do not have the same limitations). However, pragmatic clinical trials seem to outperform observational studies in terms of both epistemic rigour and practical relevance of evidence. Choices such as routine care comparators are, by definition, baked into the pragmatic design. Observational studies, by contrast, can have pragmatic features that increase the study's practical utility. Yet it is not required that they do so; all an observational study requires is the lack of randomisation. A pragmatic randomised

trial should therefore generally be preferred over a comparable observational study.

However, pragmatists have yet to theorise about pragmatic interventions and define the limits of a successful pragmatic intervention. The proposal I have developed is the idea of conceptualising these interventions as therapeutic actions, but this only points to the beginning of such a theory. In addition, the bold pragmatic attitude towards blinding as a non-essential property is most convincing if the risk of unnecessary harm can be neglected for one reason or another and if we have confidence that patients complied with the protocol to the extent required by our epistemic goals. Throughout the last two chapters, I have proposed several alternatives to blinding to achieve these goals: aligning trial designs with the interests of research participants; increase control by measuring the extent of protocol violations; increase the threshold for the minimal clinically important difference; building on prior knowledge from explanatory trials. Pragmatism can yield new insights into the effectiveness of medical interventions but it is certainly not a free pass for sloppy trial designs.

Chapter 3
The practical value of pragmatic clinical trials

Among the many criticisms levelled against randomised trials, the most persuasive is that their results do not easily apply outside the context of the experiment. The rhetoric of 'real-world evidence' cleverly exploits this concern by promising evidence about the 'real world' that applies to 'real patients'. Pragmatic trials are widely perceived as a source of evidence that is widely applicable and easily generalisable. This point has sparked the interest of bioethicists because the gain of social value is considered necessary for clinical trials to be ethical (Kalkman et al. 2017a; Borgerson 2013). The potential value of pragmatic trials has also prompted the hope of evidence that is epistemically rigorous and practically useful among clinical researchers (Thorpe et al. 2009; Ioannidis 2016; Tunis et al. 2003; Mc Cord et al. 2018; Zwarenstein and Treweek 2009; Hemkens 2018). Few scholars have thoroughly scrutinised the practical value of pragmatic trials. Kalkman and colleagues specify that many perceive that 'the pragmatic trial has social value due to the fact that it generates real world knowledge that is directly applicable to decision-making'. They analyse three different interpretations of the added social value including 'real world relevance', 'real world answers' and the probability of direct uptake of the results by decision making (Kalkman et al. 2017a, p. 140). Borgerson explains the additional value of pragmatic trials in terms of increased 'direct social value'. This is in contrast to explanatory trials, which have only 'indirect social value', as their results apply only indirectly to problems of clinical practice (Borgerson 2013). Some philosophers have suggested pragmatic trials as a potential solution to the vexing problem of extrapolation (Fuller 2019; Howick et al. 2013a; La Caze 2017). Cartwright questions whether evidence from pragmatic trials is better applicable than evidence from conventional trials (Cartwright 2017; for a general version of this argument see Cartwright 2009).

In this chapter, I substantiate three rationales that could explain the practical usefulness of pragmatic clinical trials. I begin with introducing the properties of pragmatic trials according to the PRECIS-2 tool (section 1). Then I turn towards discussing the three rationales (section 2). The first is the modest idea that pragmatic trials explore questions that are notably relevant for clinical decision-

making (section 2.1). In this rationale, evidence from pragmatic trials is practically useful simply because it provides an answer to a question that is generally of interest to clinicians and patients. This rationale could be called the 'applicability' of pragmatic trials. The second and far more ambitious idea is that pragmatic trials can support extrapolation inferences due to their *naturalness* (section 2.2). In this sense, pragmatic trials are useful practically because they can inform treatment decisions in a particular context, outside the trial. This rationale could be called the 'generalizability' of pragmatic trials. The third alternative that I discuss relates to the desideratum of causal robustness (section 2.3). It develops the intuition that routine care conditions are often deficient conditions for treatments to be effective by proposing a definition of *non-idealness.* I will show that all three have something to say about the practical and epistemic value of pragmatic trials but none of them is fully convincing. I finally defend the view that combines a modest interpretation of these proposals holding that pragmatic trials provide effect estimates that are more *realistic* than effects from explanatory trials because they are a) conducted under a set of natural conditions and b) these conditions *tend* to be non-ideal (section 2.4).

1. Pragmatic trials as natural experiments

On a methodological level, whether a trial is pragmatic or explanatory is mostly constituted by operational features (i.e., institutional, procedural and material properties) rather than high-level methodological considerations, such as randomisation and blinding. The PRECIS-2 tool by Loudon et al. conceptualises pragmatic trial designs along nine domains, including prominent design features such as eligibility criteria and the comparator treatment. It also covers less prominent features, such as the strategies to recruit patients and the level of expertise that is required to administer a treatment (Loudon et al. 2015).

In the PRECIS-2 tool, most of the nine domains are ranked according to the (informal) degree of deviance between the experimental and natural healthcare contexts. A ranking of 'very pragmatic' in the domain 'organisation' implies 'making use of no more than the existing healthcare staff and resources in that setting'; by contrast, a trial in this domain is 'very explanatory' if it relies on many such additional resources. Similarly, to rank very pragmatic in the domain 'follow-up' Loudon et al. propose to 'have no more follow-up of recipients than would be the case in usual care'. In contrast a trial will rank very explanatory in this domain

if it requires many additional follow-up visits only for research purposes (Loudon et al. 2015, pp. 6–8). It is precisely to avoid such follow-up meetings for data collection that pragmatic trials typically analyse data that is stored in electronic health records; here, the data collection process is integrated into healthcare and is indirectly made available for research. Hence, the use of routine data in pragmatic trials is not just a convenience but serves a genuine epistemic function. For these domains the contrast between *naturalness* and *artificiality* can generally distinguish between pragmatic and explanatory trials. *Naturalness* is the therapeutic situation as it would take place without the experiment and *artificiality* is every deviance from that state induced by the experiment.

Two other dimensions of the tool concern the comparator intervention and the outcome measures. For these two domains it is not so much the deviance from the routine care context but rather the clinical relevance that ranks pragmatic on the PRECIS-2 tool. For the comparator, the most pragmatic choice is to compare the intervention with 'usual care', and the outcomes should ideally be patient-relevant.

I illustrate the PRECIS-2 tool with the Salford Lung Study as an example. The study tested the effectiveness of the Relvar Ellipta inhaler to treat COPD and asthma in the area of Salford (Vestbo et al. 2016). The study implemented several pragmatic elements. For example, the general practitioners of the participants were the primary investigators in the trial, and patients were recruited at the primary care practices by their healthcare professionals. The treatment inhalers were supplied through local pharmacies. The outcome data was primarily collected using an electronic health record system by NorthWestEHealth. To enrol patients, no standardised diagnostic test or expert clinical judgment was required, instead, a COPD diagnosis by a general practitioner was sufficient. The control group was treated with the same treatment they received for their usual care, which meant the control patients took a variety of different control medications, ranging from monotherapy to triple therapy (as is usual in the treatment of COPD). Physicians were allowed to adjust the therapy of all patients to achieve optimal care. Additionally, only a few broad inclusion and exclusion criteria were employed. The primary outcome of the study was the number of severe exacerbations (acute worsening of the condition). The trial lasted one year.

The Relvar Ellipta inhaler was also tested in a trial that can be seen as its explanatory counterpart (Agustí et al. 2014). This trial compared the Relvar Ellipta inhaler to another active two-component inhaler for only 12 weeks. All participants had to take their active treatment together with a placebo to maintain blinding. The

seven inclusion criteria required a COPD diagnosis in accordance with the definition by the American Thoracic Society and the European Respiratory Society and spirometry test results that delineate COPD from asthma. The researchers additionally stated 17 exclusion criteria. Moreover, the trial contained a two-week so-called placebo run-in period to select only patients with good adherence and inhaler techniques. The primary outcome was a measure of lung function, operationalised as the amount and speed of air that could be inhaled and exhaled in a predefined time. All sites were equipped with a spirometry measuring device, and study personnel were trained in their use.[14] Table 2 compares the two trials across the nine domains of the PRECIS-2 tool in detail.

Table 2: Comparison of the Salford Lung Study on COPD across the PRECIS-2 domains

PRECIS-2 dimension	**Explanatory trial**	**Pragmatic trial**
Study identification	Title: A comparison of the efficacy and safety of once-daily fluticasone furoate/vilanterol with twice-daily fluticasone propionate/salmeterol in moderate to very severe COPD (Agustí et al. 2014) DOI: 10.1183/09031936.00054213 NCT identifier: NCT01342913 GSK Identifier: HZC113107	Title: Effectiveness of Fluticasone Furoate–Vilanterol for COPD in Clinical Practice (Vestbo et al. 2016). DOI: 10.1056/NEJMoa1608033 NCT identifier: NCT01551758
Comparator Was a clinically relevant comparator used?	fluticasone propionate/salmeterol (Inhaled glucocorticoids and LABA)	Usual care: 12%: single component therapy (LABA, a LAMA, or both) 34%: combination dual therapies (glucocorticoids, OR a combination of inhaled glucocorticoids and a LABA, OR a combination of inhaled glucocorticoids and a LAMA)

[14] Despite the overtly significant differences between the Salford Lung Study and conventional trial, the contrast could be even more accentuated. Most importantly, the explanatory study used an active comparator treatment, which is considered a pragmatic choice, whereas the most explanatory choice would be a placebo control. Moreover, Dal-Ré (2018) rated the pragmatic Salford Lung Study retrospectively using the PRECIS-2 tool and reasoned that the average score of the COPD trial was only about 2.8 (where 1 = very explanatory and 5 = very pragmatic). Their criticisms were that the recruitment was supported by a local advertising campaign and accompanied by a lengthy consent process; eligibility criteria did not perfectly match the target population, as approved by the EMA; and the package of the investigational inhaler contained a warning 'investigational drug – for clinical trials use only'.

		54%: combination triple therapy (inhaled glucocorticoids, a LABA, and a LAMA)
Eligibility Were all patients eligible who would receive the treatment in clinical care?	Seven inclusion and 17 exclusion criteria, among which are the following inclusion criteria: • Established history of COPD according to the ATS/ERS definition • Above the age of 40 • Spirometry test values with typical criteria to delineate COPD from asthma • Long-term smoker (one pack daily for more than 10 years) • Women without child-bearing potential during study (physiologically incapable or using effective contraceptives) Exclusion criteria contain: • Diagnosis of asthma or other respiratory diseases • Poorly controlled COPD that requires treatment • Carcinoma • Use of certain medication • Subjects who are unable to withhold certain other treatment 4 hours prior to spirometry testing at each visit	Five inclusion and seven exclusion criteria, among which are the following inclusion criteria: • Subjects with documented diagnosis of COPD from a general practitioner • Above the age of 40 • Current COPD maintenance therapy • Exacerbation history • Women without child-bearing potential during study (physiologically incapable or using effective contraceptives) Exclusion criteria contain: • Life-threatening condition • Unstable COPD • Chronic user of oral corticosteroids according to the opinion of the general practitioner • Subjects who plan to move away from the geographical area
Recruitment Are patients recruited using low-threshold methods?	unknown	According to (Dal-Ré 2018), recruitment was linked to an advertisement campaign and therefore not very pragmatic
Setting The choice of healthcare institution for the trial (e.g. primary care vs. specialised research facility)	unknown	General practices in the area of Salford, UK
Organisation The resources used at the healthcare institution, e.g. diagnostic devices, healthcare personnel	All outcome assessments were conducted using standardised equipment. All sites were issued with Biomedical Systems (BMS) Vitalograph 6800 Fleisch pneumotach for spirometry assessments prior to study start. Study personnel underwent training for the use of the pneumotach by BMS.	Data was captured using an electronic health record system connecting primary and secondary care by NorthWestEHealth The medication was supplied through the local pharmacies and investigators were the general practitioners. Both general practitioners and pharmacists were trained in the basics of good clinical practice
Flexibility delivery	Only several specified additional treatments were allowed and list of not	General practitioners were able to adjust medication throughout the study to

Does clinical expertise / preferences enter in the delivery of the medication or is it rigidly standardised?	permitted medications during the study exists	allow for optimal treatment of COPD, and patients were allowed to switch from FF/VI to usual care
Flexibility adherence How tightly are patients controlled in their adherence to the treatment?	The trial used a 2-week placebo run-in period to obtain baseline assessments and to evaluate adherence with study treatment and procedures, diary card completion and assessment of disease stability Compliance with treatment was assessed by reviewing the dose counters on both inhalers at randomisation (day 1), day 28, day 56 and on day 84	Used inhalers had to be returned to the pharmacies to assess adherence
Follow-up Are patients followed-up by meetings in addition to what is needed for patient care?	Unclear. Outcome assessment only at screening, day 1 and day 84	Face-to-face visits only at initiation and the end of the trial Patients were followed via the electronic health record (EHR) database; three monthly phone-calls for safety check-up
Primary outcome Were the primary outcomes chosen relevant for patients?	24-h effect after 12 weeks on lung functioning, measured as forced expiratory volume in 1 s (FEV1) – the volume of air exhaled in the first second during forced exhalation after maximal inspiration.	Severe exacerbations – i.e., an acute increase in the severity of the condition (requiring contact with the GP)

The pragmatic trial resembled the natural therapeutic situation in many ways, particularly how patients were diagnosed, treated and cared for during the trial. The explanatory trial by contrast went to great lengths to create a highly artificial research setting. Further meta-research on this case reinforces two striking differences. First, the population in the explanatory trial was considerably narrower. A meta-research study comparing the Salford Lung Study to six large conventional trials on COPD reinforces this point (Woodcock et al. 2018). Based on a retrospective analysis of a database, the authors reasoned that about half of COPD patients who were registered in a general practice in the area were eligible for the Salford Lung Study; of those eligible, about half eventually participated in the trial. This sample represents a quarter of the population. In comparison, only 15% of the same patient population would have been eligible for any of the conventional trials. The authors further found that patients in the Salford Lung Study were older, had higher exacerbation rates, a higher proportion of current smokers and a higher rate of comorbidity than the population in the conventional studies. The second difference concerns the reported adherence rate. The explanatory trial reported average adherence rates as high as 97.5%, whereas the normal rate in routine care

is expected to be well below 50% (Lareau and Yawn 2010). Presumably, artificial measures such as the placebo run-in period, the short-term follow-up for a chronic disease and several follow-up visits all contributed to the high adherence rate in the explanatory trial.

In addition, the trial includes features that are relevant in the therapeutic setting. In the Salford Lung Study the comparator intervention was usual care that included three different treatment regimens commonly administered to COPD patients. Moreover, the primary outcomes differ in their practical relevance. The Salford Lung Study measured exacerbations as primary outcome. An exacerbation is an acute worsening of the condition, which – almost by definition – severely interferes with the functioning of a patient in daily life. In contrast, the explanatory trial measured lung function as the forced expiratory volume in 1 second (FEV1). This measure is widely used for diagnostic or therapeutic decision-making, yet its clinical relevance can only be established by additional evidence. Finally, the Salford Lung study comes closer to establishing long-term effects with its length of 12 months rather than 12 weeks. It is intuitive that the pragmatic trial as a natural experiment with a focus on clinically relevant features has a unique practical value. The next section discusses what the practical value of these trials is.

2. The practical value of pragmatic trials

The methodological literature proposes some notions to further contrast pragmatic trials with their explanatory counterparts but they seldom explain what grounds the assumption about their practical usefulness. One often reads that pragmatic trials measure the 'effectiveness' rather than 'efficacy' of a treatment and that they are conducted 'under routine care conditions'; by contrast, explanatory trials are conducted 'under controlled conditions'. Pragmatic trials 'maximise external validity', while explanatory trials 'maximise internal validity', or explanatory trials answer the question 'Can it work?' while pragmatic trials answer the question 'Does it work?' (Luce et al. 2010); for an overview of these notions, see Table 3. Particularly, the intuition that pragmatic trials increase external validity at the expense of internal validity is commonly cited to explain the epistemic merits, problems and tensions of these designs (Godwin et al. 2003). While these ideas offer important explanations, they also introduce potential misunderstandings and new puzzles. An example is the question of how internal and external validity relate

to each other (Jimenez-Buedo and Miller 2010; H. Chytilová and R. Maialeh 2015). I develop three rationales that could explain the practical usefulness of these trials and show how they relate to these notions.

Table 3: Methodological comparison between pragmatic and explanatory trials

	Conventional trial	**Pragmatic trial**
Validity	Internal validity	External validity
Causal conclusion	Efficacy	Effectiveness
Epistemic aim	Explanatory	Pragmatic
Causal question	Can it work?	Does it work?
Background conditions	Experimental Artificial Homogeneous	Routine care Natural Heterogeneous

2.1. Applicability to decision-oriented research questions

Pragmatic trials are practically valuable in the modest sense that they ask a question that is relevant for decision-makers in healthcare. A typical question in a pragmatic trial mirrors the logic of decision-making in healthcare by measuring patient-relevant outcomes and choosing a practically relevant comparator. Unlike many traditional randomised trials, pragmatic trials encourage 'usual care' as comparator, which means comparing the new intervention with whatever care patients would receive without the experiment. Hence, the comparative question in the trial would be 'Is A more beneficial than B?' Here, A and B both represent relevant options to choose from, and the question directly mirrors the logic of the clinical judgement of choosing between different treatment options. A second reason why pragmatic trials fill decision-relevant knowledge gaps is that they encourage the use of patient-relevant outcomes. Such outcomes are relevant to inform patients' choices because they measure effects that matter to patients. In contrast, trials that measure surrogate outcomes or outcomes on the physiological level are less relevant for a patient's decision about treatment choices, because their clinical benefit or patient-relevant benefit is uncertain (Fleming and DeMets 1996). In addition, pragmatic trials can relatively easily measure long-term outcomes, particularly if they use data from an electronic health database. Such trials have reportedly measured long-term outcomes with zero loss to follow-up (i.e., the problem of investigators losing track of patients after a while) (Hemkens 2018). In contrast, explanatory trials

demanding high resource investment and compliance rarely maintain an acceptable follow-up rate for a long period; they risk patients dropping out because of the burdensome requirements for participating in the trial. This is the modest sense that Borgerson argues that pragmatic trials have 'direct social value' while explanatory trials have 'indirect social value' (Borgerson 2013). Kalkman and colleagues also highlight these benefits of pragmatic trials to substantiate the view that pragmatic trials have 'real world relevance', and provide 'real world answers' (Kalkman et al. 2017a).

Another potentially interesting feature for decision-making is pragmatic interventions. Chapter 2 showed that the medical interventions in pragmatic and explanatory trials differ substantially. Explanatory trials allow an inference about the precise contribution of the pharmacological properties of the drug that are involved in the drug's mechanism of action. For this reason, Schwartz and Lellouch use the term 'explanatory' trial, because these trials contribute to our mechanistic knowledge, which can explain *why* drugs are effective and not only *that* they are effective. Pragmatic trials only support inferences to broader intervention notions that can include contextual causal factors like adherence or awareness about the medicine. One advantage of such interventions is that they can make the impact of contextual factors visible. I discussed this aspect in Chapter 2 and the example of the Relvar Ellipta inhaler illustrates this point. Particularly in situations where physicians and patients can choose between different drugs of the same class with similar effectiveness in controlled trials, the choice for the most effective therapeutic action is plausibly determined by other properties of the medicine. If so, evidence that can make the impact of such causal factors visible is clearly practically useful.

According to some pragmatists such pragmatic interventions have also increased applicability. Zwarenstein and colleagues argue that 'Pragmatic trials may incorporate these factors into the estimate of effectiveness, rendering the findings more applicable to usual care settings' (Zwarenstein et al. 2008, p. 6). The idea that pragmatic interventions are more applicable seems counterintuitive because these interventions contain causal factors – such as patient preferences or physician skill levels – that are highly contingent and change across settings. Hence, implementing them in settings outside the trial would also require bringing about all these contingent factors. This point is true, but the same holds for traditional medical interventions. An explanatory trial standardises causal background factors so as to measure the treatment effect independently from the influence of these contextual factors. Yet, this does not mean that the medicine is effective regardless

of the contextual factors. We still require patients to adhere to the treatment, physicians to administer the treatment correctly and so forth. The experimental intervention that brings about the effect in an explanatory trial not only administers the medicine but also includes the training of physicians or the monitoring of patients. Bringing about the effect elsewhere would require physicians to get trained, administer the treatment and monitor patients. The pragmatic experiment, only intervenes to the degree that also a physician can do in healthcare. Doctors can often only prescribe, recommend or administer a treatment. Hence, the effect estimates from a pragmatic trial are practically valuable for treatment decisions because they provide an estimate about the kind of interventions that physicians can make.

The first proposal suggests that pragmatic trials are practically valuable and applicable to decision-making in the sense that they ask questions, measure outcomes and implement interventions that closely mirror practical interests. Such evidence is relevant to health care professionals because it provides an answer to question that is relevant for healthcare professionals. This is the modest sense in which pragmatic trials can be said to be *applicable* to practical contexts. Although I acknowledge that this contribution is already valuable, advocates of these trials clearly make stronger claims when they hold that these trials are highly applicable and generalisable. Furthermore, most of the dimensions of pragmatic trials in the PRECIS-2 tool are not required for this benefit. Researchers can measure outcomes and comparators that matter to healthcare professionals and patients in clinical trials that are purely conventional in all other ways. If this is all we can gain from pragmatic trials it seems that we do not need the explanatory-pragmatic distinction to create such benefits.

2.2. Pragmatic trials as a solution to the problem of extrapolation

A prevalent idea about pragmatic trials is that they are highly generalisable or have high 'external validity' and have therefore been proposed as a response to the problem of extrapolation by several philosophers. (Fuller 2019; Howick et al. 2013a; La Caze 2017). Problems of extrapolation consist in the challenge that we have some evidence about a causal relation in a study population and want to infer whether the relation will hold in another population of interest, i.e., the target population. Thus, the general problem of extrapolation in clinical research is concerned with inferences of this sort:

(1) C causes E in Context1

from which we would infer that:

(2) C will cause E in Context2,

where C is the cause, E the effect and $Context_i$ is the context in which the causal relation holds. The problem of extrapolation is concerned with the question of what kind of evidence and reasoning strategies would allow us to infer (2) from (1). Strategies to deal with the problem of extrapolation mostly attempt to establish sufficient similarity between the context or populations in the study and that of the target. Researchers do this by establishing that both contexts are governed by analogous biological and social mechanisms (See for example Howick et al. 2013b; Steel 2008; Guala 2010; Cartwright 2009). The complexity of these philosophical proposals is in stark contrast with the far simpler strategies used in clinical research, and philosophers have repeatedly argued that these simple strategies are an unreliable guide to extrapolation (Fuller 2013, 2019; Stegenga 2018; Cartwright 2017; Howick et al. 2013b). However, many of the arguments against these simple strategies are based on the assumption that all clinical trials are explanatory trials. The question arises to what extent simple extrapolation strategies can be a reliable guide if they draw on results from pragmatic clinical studies from the outset.

I briefly summarise the three simple strategies. The first strategy advocated by proponents of the EBM movement in clinical research has been called simple induction or simple extrapolation. As its name suggests, this simple solution implies that results from standard clinical trials are 'generally generalisable'. Fuller argues that neither the empirical evidence nor arguments from theory support the general generalisability thesis (Fuller 2013, 2019). Empirical evidence in support of this thesis is provided by converging results from RCTs; yet these results are unfaithfully selected, as RCTs often do not converge. In addition, evidence from within multiple RCTs does not extrapolate beyond the RCTs (Fuller 2013). In his 2019 article, Fuller argues further that mathematical theory does not support the assumption that the usual outcome measure is insensitive to differences in causal factors. Furthermore, we cannot conclude from biomedical theory that there are no such differences of causal factors (Fuller 2019). Indeed, as Stegenga points out, the opposite is true, as the medical community knows that environmental or physiological differences can modify treatment effects (Stegenga 2018, chapter 8). For example, there is ample evidence that the expertise of a surgeon, the level of infrastructure at a care site and the timing of a treatment can be crucial to treatment success.

Another slightly more sophisticated strategy is what Stegenga calls 'simple-extrapolation-unless' (Stegenga 2018, chapter 8). The strategy is similar to simple extrapolation but adds an 'unless clause', which states that the general generalisability principle holds *unless* there is a compelling reason why it might not. Fuller argues that proponents of this thesis fail to show what would count as a compelling reason; hence, the thesis collapses into the general generalisability thesis. A second objection concerns the logical fallacy of the clause itself. The unless clause, Fuller argues, is based on an argument from ignorance, which concludes from the fact that one does not know that P to not P. Such arguments are only reasonable if the knowledge base that would include P, if P were true, is reasonably complete. This is a profoundly mistaken assumption in the case of medicine (Fuller 2019). This intuitive point is reinforced by Stegenga, who emphasised that a vast evidence base is withheld from the medical community because investigators fail to make mostly negative trial results public; this is called the problem of publication bias (Stegenga 2018, chapter 8).

A third strategy to cope with the problem of extrapolation relies on the inclusion and exclusion criteria to identify the subpopulations to which the results are generalisable. If some individuals who belong to a certain subpopulation participate in a study, then (so the reasoning goes) the result is generalisable to all individuals of that subpopulation. Cartwright pointed out several problems with this line of reasoning. First, the inclusion and exclusion criteria to delineate subpopulations are usually based on criteria such as gender, ethnicities, stages of disease, age and so on (Cartwright 2012). They constitute a mixture of biological, social and clinically relevant criteria – a 'potpourri' in Cartwright's words, lacking systematicity. In addition, such criteria are only an indirect indicator for the relevant causal mechanisms. The criteria can only serve as the basis for extrapolation if they delineate the subpopulations such that the relevant causal mechanism is shared among all the individuals fulfilling the criteria and no other mechanism is present that overrides the treatment effect. It is doubtful whether these conditions are ever satisfied by superficial eligibility criteria.

Although pragmatic researchers cannot respond to all the objections raised against these simple extrapolation strategies, it is intuitive that pragmatic trials (which resemble the natural therapeutic environment) are good candidates to get them off the ground. However, the simple strategies still do not get us very far. To gain practical value, we require at least that the results extrapolate in time; i.e., results extrapolate to future patient populations within the same setting. Such an extrapolation in time holds if two conditions are met:

1) The trial population is representative of future patient populations within the scope of recruitment.

2) The therapeutic context is sufficiently stable across time.

A well-conducted pragmatic trial has a good chance of fulfilling condition 1) to a sufficient degree; indeed, that is why the design is generally valued. The merit is commonly attributed to the few eligibility criteria of pragmatic trials; however, the hard epistemic work is done by the low-threshold recruitment techniques of patients. Eligibility criteria only define who can and cannot participate. The correct recruitment techniques ensure that the identified patients in fact participate in the trial. Explanatory researchers go to great lengths to recruit the ideal patients into their studies, for example, by sending out invitation letters with questionnaires. Such high-threshold recruitment techniques introduce considerable biases regardless of how many exclusion criteria are explicitly mentioned. Pragmatic trials on the other hand employ recruitment in an inclusive, low-threshold manner, which usually means that patients are invited to participate in the trial when they contact the healthcare facility for a routine visit. Such strategies ensure that all patients who would receive a new treatment in routine care – within the scope and period of the trial recruitment – are *offered* participation in the trial. The result of the pragmatic recruitment efforts is what we might call an *inclusive* patient population. The inclusiveness speaks primarily in favour of the method's ethical value, as it contributes to justice in health research. However, inclusivity can also speak towards representativity if three requirements hold: First, we need to assume that a mostly unbiased proportion of the patients who are invited to participate also consent to participate. Arguably there might still be some self-selection bias at this stage – which has prompted discussion about whether it is justifiable for pragmatic trials to overlook the informed consent procedure to mitigate this problem (Kalkman et al. 2017b). Despite these difficulties, I think we can be generous and assume that the remaining bias is acceptable. Second, if the target group is described as 'all patients with COPD in the Salford area', pragmatic recruitment techniques would usually overrepresent frequent healthcare users. To mitigate this problem, researchers should adopt a pragmatic definition of the target population, such as 'all patients who would receive the new treatment in routine care'. Since frequent healthcare users are more common in both the trial and the target population, bias is avoided. Third, the period of the recruitment must be comparable to future periods. For example, patients participating in a trial during a public health crisis might not be representative of future patient populations within the same scope of recruitment.

What about condition 2)? Because pragmatic trials preserve the natural therapeutic environment, the sameness of contexts (i.e., between the experimental and therapeutic environments) can be assumed as long as the context remains sufficiently stable. Assessing the stability of the healthcare context requires background knowledge about causally relevant contextual factors. However, such an assessment only requires assessing the causal relevance of changing factors in the therapeutic environment, rather than establishing the similarity among all causally relevant factors across two contexts. Hence, establishing stability within a context is epistemically less demanding than establishing the similarity between contexts.

To summarise, results from pragmatic trials have a good chance to extrapolate in time within the scope of recruitment *until* there are significant changes that interfere with the stability of the healthcare context or the composition of the patient population. A modified version of the simple strategy of 'extrapolation unless' applies and involves what we could call 'simple extrapolation until'; here, significant change in the context would prevent the simple extrapolation. If the few epistemically relatively undemanding assumptions outlined above hold, extrapolation in time is justified within the scope of recruitment and for a limited period in the history of healthcare. Hence, they respond to some local versions of the extrapolation problem.

While extrapolation in time comes relatively cheap, extrapolation in space is a different matter. The naturalness of an experiment warrants its approximate sameness within the setting and scope of recruitment, yet this property cannot do much work beyond this context. Inclusive sampling cannot speak in favour of the representativity of the study population if the recruitment involves a patient population that differs from the target population in relevant respects. The pragmatic trial by Vestbo et al. on COPD, for example, was conducted in Salford, a region that was highly affected by the post-industrial crisis and thus presents a peculiar research context – to what extent then are its results applicable to patients in Swiss hospitals? The study population in the study was sufficiently representative of future patients in the Salford area. But it is probably not a representative sample for regions that display different demographics, social-economic status, smoking policies, healthcare provision or air quality. Extrapolation in space not only requires assessing the stability of one healthcare context but also the similarity *between* contexts, which is epistemically far more demanding. The former only requires assessing the causal relevance of changing causal factors, the latter requires comparative assessments regarding all causally relevant factors, and these are many.

For example, healthcare settings in different countries are equipped and staffed differently, they use different standards of care or different insurance policies might determine what treatments are administered. Such similarity judgements about relevant causal factors require vast evidence and background knowledge. There is little that pragmatic trials can do to meet these requirements.

Unlike the simple strategies recommended by the EBM movement, pragmatists are painstakingly aware of the limitations for generalising the results of pragmatic trials across contexts. For this reason, the CONSORT extension for the transparent reporting of pragmatic trials requires, first, the reporting of key causal factors in the local setting. Second, researchers must 'discuss possible differences in other settings where clinical traditions, health service organisation, staffing, or resources may vary from those of the trial' (Zwarenstein et al. 2008, p. 6). This recommendation does not amount to a simple extrapolation strategy, because such an assessment again needs substantial evidence from outside the trial. The second point is in line with Cartwright's argument that the limits of applicability of pragmatic trials are approximately the same as the limits of applicability for ideal trials; any attempt to simply extrapolate pragmatic trial results across contexts does not get us very far very easily. There is, however, a crucial difference between the two trials. The naturalistic character of pragmatic studies at least nourishes the hope that searching for other settings with similar conditions is not in vain. Conversely, such hope is indeed misplaced for most explanatory trials, as their artificial experimental conditions almost certainly do not occur anywhere else[15]. Hence, we should keep in mind that explanatory trials neither extrapolate in time to the context in which it has been conducted because this context ceases to exist after the trial ends, nor is it plausible that the conditions of an explanatory trial can be found in any other natural health care setting. Nevertheless, the pragmatic trial's contribution to issues of extrapolation is relatively modest, and these trials cannot fulfil the promise of their great generalisability in virtue of their naturalness.

Pragmatic trials have been criticised not only for not solving the problem of extrapolation but even for making the problem worse (Cartwright 2017; La Caze 2017; Gedeborg et al. 2019). I consider this criticism briefly. Cartwright and others repeatedly and convincingly argued that an average result from a clinical

[15] Many hold that the most promising route to extrapolation is by mechanistic knowledge. The reason why Schwarz and Lellouch opted for the term explanatory trial is because, in their view, these trials contribute to our mechanistic knowledge about the causal mechanism of action by measuring narrow interventions and physiological outcomes. From this perspective, explanatory trials may well outperform pragmatic trials to solve extrapolation problems.

trial only establishes that there is *some* set of conditions in which the causal relation holds; i.e., 'it works somewhere'. A group trial can neither establish that the causal relation holds under all conditions included in the trial nor that it holds under any particular set of conditions included in the trial. Indeed, a positive average effect can be perfectly consistent with two opposite effects for two different subpopulations. For these reasons, it has been argued that the problem of extrapolation begins *within* randomised trials and not only outside them. In La Caze's words, it is 'the main selling point' of pragmatic trials to increase heterogeneity, which in turn exaggerates rather than solves the problem of extrapolation (La Caze 2017). Karnicoloas and colleagues argue against the pragmatic-explanatory distinction on related considerations. They hold that evidence from a trial that include non-compliant patients does not have any bearing on the outcomes for highly motivated patients (Karanicolas et al. 2009) . Consequently, the solution preferred by Cartwright and others who follow this critique is to abandon the project of averaging treatment effects over heterogeneous populations and instead investigating the causal efficacy of potential effect modifiers individually (Cartwright 2017; Gedeborg et al. 2019).

The problem of subgroups within clinical trials prompts a qualification to the above discussion. First, pragmatic trials, just like any other group-comparative method, support only population-level causal claims and, consequently, population-level extrapolation inferences. If one is looking for evidence to predict outcomes for individuals, Cartwright is right that any clinical trial – whether explanatory or pragmatic – is not a suitable source of information. Hence, it is only at the population level that pragmatic trials yield the modest advantage for extrapolation of population-level average treatment effects discussed above. A related concern is that such population-level average treatment effects are practically meaningless. Gedeborg illustrates the concern by arguing that it would be practically meaningless to average the treatment effects across adult and paediatric populations in a pragmatic trial (Gedeborg et al. 2019). However, in situations where clinical practitioners do not differentiate between subpopulations in the trial because the causes of treatment variability are unknown, the average treatment effect is probabilistically informative for making treatment decisions about future patients. This does not mean that the search for the causes of treatment variability can be abandoned. Still, it seems worth reinforcing that inclusive patient populations in pragmatic trials do not justify an inference to any particular individual of a certain subpopulation included in the trial and, a fortiori, effects cannot be generalised to all individuals of these subpopulations outside the

trial. Hence, efforts to specifically include marginalised and otherwise under-represented communities in clinical trials might be epistemically overrated by the medical community. (Nonetheless, I think these efforts are worth pursuing for reasons of health justice.)

I do not think that the problem undermines the epistemic value of pragmatic trials. However, it reinforces the concern that pragmatic trials have at best only a moderate advantage in terms of addressing extrapolation problems through their naturalness. In the Annex to this thesis I propose a definition of the notion of external validity that captures this moderate epistemic advantage of experiments to respond to extrapolation problems.

2.3. The epistemic value of non-ideal trials

I develop a third proposal that might explain the value of pragmatic trials, namely, the ability of these trials to provide robust conclusions about effectiveness. In this section, I introduce the ideal vs. non-ideal distinction to show how non-ideal trials can support robust positive causal conclusions. The distinction captures the intuition that pragmatic trials are conducted in unfavourable conditions that generally reduce the effectiveness of medical intervention rather than increasing it. If this intuition is true (or at least we could identify cases where it is true), this rationale could be of considerable epistemic value in support of the pragmatic approach.

Rather than focusing on the role of naturalness in pragmatic trials, I ask what purpose artificiality serves in clinical trials. I have so far discussed two roles of control measures, namely controlling for biases and increasing the homogeneity in a trial. Both fail to account for all standardisation practices in explanatory trials. Notably, homogeneity cannot explain why clinical researchers commonly do not merely standardise background conditions at an arbitrary value but rather employ specific values. For instance, a clinical trial generally does not standardise adherence at 50% – which would satisfy the need for homogeneity – but rather at the rate that is *ideal* for the treatment under scrutiny. To explain this practice, I draw on the notion of an INUS condition. The acronym stands for 'insufficient but necessary part of an unnecessary but sufficient' condition. Medical treatments are thought to be such INUS conditions, meaning that the medicine is a necessary part for the effect to occur but is by itself insufficient. In addition to the treatment, numerous other conditions must be in place, such as accurate patient diagnosis, adherence to the treatment and the right genetic make-up for a response to treatment. The primary objective of many artificial controls is to meticulously

guarantee the presence of all other causal factors to maximise the effectiveness of the treatment. Hence, what is peculiar and epistemically pertinent about explanatory trials is that they artificially interfere in the causal setup not only to standardise but also to *idealise* many INUS conditions. That is, they employ an *ideal* set of values – such as ideal age, ideal disease states, ideal adherence or ideal treatment conditions. For example, explanatory trials exclude patients with comorbidities because those patients often respond less effectively to treatments. Furthermore, patients are motivated to adhere to a treatment at a rate that is known to be ideal for the treatment to be effective or treatments are administered in a way that is known to be ideal for the treatment to be effective. Due to these ideal background conditions, explanatory trials are epistemically privileged in the sense that they increase our chances of observing a causal effect if one exists.

'Idealness' here refers to the values of supporting causal factors in the causal complex. Determining an ideal value is a purely empirical question. Assumptions about what is ideal adherence or ideal dosage are built on prior empirical evidence from drug development programmes or other evidence. Such a notion of an ideal trial is distinct from the other notions of ideal trials discussed by Cartwright (2009) and Reiss (2019), who state that ideal trials are those that fulfil all the conditions for the causal inference to follow deductively. Consequently, determining ideal conditions in the sense introduced here is dependent on the state of knowledge about what the ideal background conditions are and the methods at hand to instantiate those conditions. Idealness in that sense is clearly an idealised assumption. Nonetheless, the assumption carries explanatory and prescriptive value, as I show in the next paragraph.

Let us define 'ideal conditions' as the values of all causal background factors such that an effective treatment is most effective. An ideal trial in that sense then is a trial that attempts to standardise background conditions at ideal values such that an effective treatment would be most effective. From this definition of ideal conditions, the following two claims can be derived:

1.1 If a treatment is effective under ideal conditions, it may or may not be effective under non-ideal conditions.

1.2 If a treatment is effective under non-ideal conditions, a fortiori, it is effective under all more-ideal conditions.

Claim 1.1) confirms what many have criticised about ideal clinical trials, namely that a positive result in an ideal trial only strictly confirms the causal conclusion

under the conditions of the trial and does not extrapolate beyond. Learning that a treatment is effective under ideal conditions does not tell us anything about how that treatment would perform under non-ideal conditions. We have only tested the treatment under the conditions under which the treatment would be most effective. Claim 1.2) is more interesting. It states that positive results from non-ideal trials generalise to all more-ideal conditions. Thus, a treatment that is effective under non-ideal conditions will certainly be effective under more-ideal conditions (because ideal conditions are those under which an effective treatment is the most effective). In other words, the fact that the positive effect does not disappear despite non-ideal conditions is evidence for the claim that the set of background conditions under which the treatment is effective is wider than the set of non-ideal conditions in the trial, including at least all more-ideal conditions. In this sense, non-ideal trials can establish more general or robust causal conclusions than ideal trials – and such knowledge is valuable practically. Its utility value lies in evidential support for the claim that the treatment's effect extends beyond the background conditions included in the trial, providing at least some supportive evidence for using the treatment outside the context of the trial.

The distinction between ideal and non-ideal conditions also has interesting implications for negative results. Again, given the definition of ideal conditions as the conditions in which an effective treatment is *most* effective, we can infer the following:

2.1 If a treatment is ineffective under ideal conditions, a fortiori, it is ineffective under all (non-ideal) conditions.

2.2 If a treatment is ineffective under non-ideal conditions, it may or may not be effective under ideal conditions.

Claim 2.2. articulates a familiar principle to proponents of pragmatic trials: If a treatment is ineffective under non-ideal conditions, it remains uncertain whether the treatment would have been effective under ideal conditions. This point holds significance. Without knowing whether treatments work under ideal conditions, we cannot discern whether the treatment is truly inert or the experimental conditions deviated too greatly from the ideal conditions. Hence, negative results in non-ideal trials are crucially underdetermined. At the same time, from claim 2.1. it follows that this is precisely the epistemic advantage of ideal trials, as their negative results generalise to all non-ideal conditions. If a treatment is ineffective under the conditions under which it would be most effective, there is no doubt that there are no conditions under which the treatment would be effective.

In that regard, ideal trials are a very efficient method to learn from negative results and eliminate ineffective treatments. This notion is equivalent to the idea that such trials can provide a 'proof of concept'; the same idea is described in terms of explanatory trials answering the question 'Can it work?' What my proposal adds to these views is that we also gain important information from a negative result, because we have certainty that the treatment does not work elsewhere.

This rationale has considerable explanatory value in support of non-ideal trials. In my view, it is a trial's non-idealness and not its naturalness that can bring support to the causal conclusion beyond the context of the trial. Its shortcoming is that the definitions of ideal and non-ideal trials only partially capture real explanatory and pragmatic trials for several reasons. First, explanatory trials are rarely truly ideal trials. Researchers rarely know who the ideally responding patients are, they cannot force patients to fully adhere to treatments and a negative result could be the result of a chance process. A single negative explanatory trial is not fully convincing to eliminate a medicine as truly inert. However, idealness can serve as an epistemic aim pursued by researchers for designing trials with the aim of maximising certainty about negative results. A second qualification to consider is that the definition of ideal and non-ideal conditions does not account for the comparative nature of clinical trials. Effect size estimates in comparative experiments depend fundamentally on the effectiveness of the comparator treatment. Even if all INUS conditions are ideally chosen, an experiment does not result in a large effect size if the treatment is compared to an equally effective alternative. The use of placebo controls is imperative to identify inert treatments. The third and perhaps most detrimental short coming is that the definition of ideal and non-ideal conditions only ranges over support factors. It neglects the role of other causal factors that affect the treatment effect independent of support factors like patient's awareness of the treatment or the intake of other medicines. Placebo effects or other uncontrolled causal factors can produce positive result in a pragmatic trial even if the treatment is ineffective under (non)-ideal conditions. The rationale from the non-idealness is only convincing to the degree that the positive contribution of such factors can be neglected. Let's take stock.

2.4.Pragmatic trials provide realistic treatment effects

I have discussed three epistemic rationales to ground the widely hold assumption that pragmatic trials are somehow better applicable, widely generalisable or have high practical relevance. The notion that pragmatic trials are practically useful because they compare treatments, measure outcomes that are relevant for

healthcare professionals is most convincing. However, the pragmatic-explanatory distinction becomes superfluent. Pragmatic trials only partially fulfil the requirements of two more demanding proposals. If the few epistemically relatively undemanding assumptions hold, pragmatic trials can solve some local version of the extrapolation problem. However, the pragmatic trial's contribution to problems of extrapolation is relatively modest. The intuition that pragmatic trials can establish more robust causal effectiveness claims because of their non-idealness is intriguing, but it also suffers from several shortcomings that it fails to be fully convincing.

Nevertheless, both interpretations have something to say in favour of pragmatic trials, but the discussions show that these benefits need to be formulated more modestly. My analysis clearly supports the critics who argue that explanatory trials are insufficient to support the use of treatments outside artificial experimental conditions and insufficient evidence for regulatory approval. Contrary to a widely hold view, I have argued that the main problem is not their lack of representativity but their tendency to be conducted under ideal conditions. These causal conclusions are particularly fragile. I think pragmatic trials go some way of addressing this problem by providing treatment estimates that are more *realistic* than those from highly controlled trials. *Realistic* I have in mind a combination of the two requirements discussed above but modestly interpreted: the conditions of pragmatic trials are realistic in the sense that they are a) conducted in a set of *natural* conditions and b) that tend to *include non-ideal* conditions.

Such *realistic* treatment effects do not simply generalise beyond the local context of the experiment, and they are not robust in the sense proposed above. Estimates about such treatment effects provide a different perspective on the effectiveness of medical interventions. This perspective is valuable because it complements effect estimates from explanatory trials that tend to provide estimates under ideal conditions.

Part II
Data pluralism

Chapter 4
Data: Good, bad or good enough

This second part of this thesis is devoted to data. In the philosophy of medicine, much attention has historically been payed to the examination and assessment of methodological attributes, such as randomisation or blinding. With the advent of real-world evidence, it becomes evident that stakeholders in clinical research are not only concerned with the plurality of research methods but also, and perhaps more relevantly, with the plurality of data. Over the past three decades, the collection and management of data in clinical trials have been meticulously governed by a comprehensive sixty-page document known as 'Good Clinical Practice' (GCP). These globally accepted rules and principles play a pivotal role in ensuring the integrity of data and the safety of patients participating in clinical trials by defining essential requirements for data collection, storage, and verification. Central to the rules of GCP is the fundamental premise that data and its purpose are inextricably linked in the sense that data are produced by data users with a specific scientific purpose in mind. Contrary to this, the new real-world evidence paradigm allows that data can be used for research even though the data were produced by someone else than their user and for an entirely different purpose. Such data are shaped by the local needs, constraints, and incentives of local health care agents rather than concerns for their epistemic value.

Sabina Leonelli, who pioneered the philosophy of data-centric science, describes such data as embarking on a *journey* (Leonelli 2016). She drew philosophers' attention to the coming avalanche of travelling data and the frictions and costs that accompany it. She holds that 'there is nothing smooth about data journeys' and that 'journeys of data require planning, involve different material infrastructure, and are generally fragmented and complex just like human journeys' (Leonelli 2016, chapter 1). As such, data journeys require laborious and skilled work to *decontextualise* data from their original purpose, to *package* them with contextual information for travelling, and to *recontextualise* them to be used for a

new purpose. Leonelli further discusses the essential epistemic work that data infrastructures contribute to data journeys, the emergence of the professional role of *data curators*, and organisational structures like research *consortia*, that regulate data journeys (Leonelli 2016, chapter 2). Consequently, data became a valuable scientific product in its own terms, which is why Leonelli refers to data-*centric*, rather than data-*driven* science. In Leonelli's framework, real-world data are data that *travel* from sources like clinical information systems to new places where they can serve as evidence for scientific claims. It inspired new perspectives on data sharing practices across disciplines (Leonelli and Tempini 2020). My study focuses on epistemic issues of data *users* and those who rely on knowledge that was generated with repurposed health data.

Data that fits Leonelli's framework is described by a range of different terms, frequently used are 'real-world data', 'secondary use data' and 'routine data'. The term 'real-world data' became the most popular among stakeholders in the field. It is a clever rhetorical manoeuvre as it turns the main weakness of this type of data into one of its most compelling advantages: The problem that such data are shaped by the practices, needs, and incentives of ordinary healthcare settings is reframed as a representational value about natural settings in general. A brief philosophical look into the notion of data illustrates what is problematic about this. A common conception in the philosophy of science is to think of data as records or traces that *represent* the world. Practices of data journeys have complicated this picture because these practices imply that data do not have fixed representational value but instead can be interpreted by various experts for yet unknown research questions. Consequently, Leonelli has proposed to adopt a *relational* view of data, where data are constituted by their ability to circulate among individuals that treat data as *potential* evidence for a scientific claim (Leonelli 2016, chapter 3). Recent theories of data restored the representational view while incorporating Leonelli's lesson that data does not have fixed evidential value. Pietsch critically discusses several definitions of data and defends a *causal representational* view of data where data represents the world in virtue of them being caused by singular facts in the world (Pietsch 2021). Bokulich and Parker similarly reconcile both aspects by proposing a *pragmatic representational* theory of data and data models. In their view data have open evidential value but their representational content is still constrained 'by the fact that they are the product of a particular set of causal factors and not others' (Bokulich and Parker 2021, p. 8). Their view is *pragmatic* because they emphasise that the evidential value of data is not only determined by how well data (and data models) represent the world but also by other pragmatic factors. These

representational proposals illustrate why real-world data are epistemically rich and poor at the same time: data that were generated in natural therapeutic settings represent therapeutic facts from these settings. Yet, such data are also causally affected by all kinds of factors like the needs or interests of healthcare professionals, reimbursement incentives, or institutional structures and policies that make interpreting these data extremely difficult. I comment on the most common problems with such data in section 2 of this chapter.

Another notion that is widely used is that of 'secondary use' data. The notion refers to the fact that such data were initially generated for a different (non-epistemic) primary purpose of a primary user. Its meaning acknowledges the data's weakness that they are not tailored to the needs of its secondary user. Researchers' needs for data generally differ greatly from the needs of those who generate the data, as researchers require exceptionally extensive, finely grained, highly accurate, complete, and representative data. Because primary users of 'secondary use data' are not driven by such considerations, the quality of secondary use data emerges as a major epistemic concern. The notion also emphasises another valuable aspect of these data as a (presumably) cheap by-product of data collection processes in healthcare. The concept of 'secondary use' data echoes with a larger vision in which health data are seen as a sustainable resource capable of circulating within a vast and interconnected network of data users and producers. Leonelli noted that the mobility of data is one of the great promises of big and open data because people with diverse expertise can interpret data and generate scientific insight (Leonelli 2020) . 'Secondary use data' highlights the aspect that data can even circulate between epistemic and non-epistemic users to make the most out of limited resources. However, there is an enormous tension between the promise of 'secondary use data' and extensive political efforts to increase data quality 'at the source' to make secondary use possible. I elaborate on this tension in Chapter 5.

Together with the advent of traveling data came the essential question of how to evaluate the quality of such data for reuse. Stakeholders in clinical research turned towards a contextualised approach to data quality. In such an approach, favourably assessed by Leonelli and Canali (Canali 2020; Leonelli 2017b), data cannot be said to be good or bad independent of a specific context of use. Bokulich and Parker developed an adequacy-for-purpose approach to data models in analogy to such an approach to scientific models (Bokulich and Parker 2021). A recurring topic of the second part of this thesis is the comparison of such a contextualised approach with the current rule-based reference standard for *good* data called 'Good Clinical Practice' (GCP). I show that this standard is a widely

neglected source of the epistemic value of conventional clinical trials to ensure not only the reliability but also the trustworthiness of data.

Data quality is known to be a notoriously multifaceted notion. Bridging the gap between the abstract idea of 'fitness-for-purpose' and concrete measures to quantify data quality is a complex task with countless possibilities to approach it (Illari 2014). Leonelli noted that the wide variability of these practices makes it difficult to define international standards (Leonelli 2017b). The thesis of the underdetermination of evidential significance articulated by Stegenga for assessing the quality of methods is strongly accentuated in practices that assess the quality of data (Stegenga 2018, chapter 7). Nonetheless, regulators are currently developing their own data quality frameworks to coordinate data quality assessments in clinical research and establish new rules for what counts as acceptable data in the field (European Medicines Agency 2022; US Food and Drug Administration 2021b). However, my investigations in Chapter 6 reinforce the worry that the deep local contextuality of these practices pose a remarkable challenge to the trustworthiness of data.

In this chapter, I introduce and contrast the reference standard for data quality, GCP, with the emerging contextualised approach, fitness-for-purpose. I establish GCP as the main reference standard for what counts as acceptable data in the clinical research community and argue that these rules contribute epistemic value to clinical trials that is widely neglected by philosophers of science. I then briefly introduce the main problems of data quality that researchers encounter with real-world data and point towards two major transformations of research and health care that these problems will bring along. First, studies will move towards more hedged clinical research questions, broader patient populations and more clinically relevant outcomes because routine data would otherwise be unavailable. Second, we will witness continuous efforts to increase data quality 'at the source' by systematic attempts to train and incentivise data collection in healthcare according to the needs of research. I then turn towards the new paradigm for data quality, 'fitness-for-purpose'. Conceptual reflection of the fitness-for-purpose approach together with a detailed case study support the view that the turn towards a contextualised approach is a two-sided sword: If used with sufficient skills, expertise and the right moral attitudes it is a powerful tool that allows the research community to carefully tailor the choice of methods and data to the community's epistemic and practical aims. At the same time the contextualised approach is highly susceptible to errors and misuse. Chapter 5 reinforces this view by introducing the complexity of data quality assessments and the difficulties to perform them reliably.

In Chapter 6 I extent this line of thinking by showing that data quality assessments raise concerns about the trustworthiness of data because they are nearly impossible to verify by impartial third parties.

Section 1 introduces the Guideline for Good Clinical Practice and the essential role of *monitors* to verify and increase the accuracy of data. Section 2 turns towards bad data and provides an overview of the common problems associated with real-world data and its broad practical and epistemic implications. Section 3 introduces analyses the notion of fitness-for-purpose and exposes the theoretical and normative commitments required to put the notion to use. Section 4 illustrates the fitness-for-purpose approach with a detailed case study about the approval of Prograf as an immunosuppressant for lung transplant recipients. Section 5 articulates the idea that a contextualised approach to data quality can be a powerful tool to increase epistemic opportunities while at the same time a free pass for misuse with little tolerance for errors.

1. Good data: Good clinical practice

The Guideline for Good Clinical Practice (GCP) published in 1997 by the International Council for Harmonisation (ICH) is a pivotal document that sets forth globally accepted conventions and principles for conducting clinical studies (International Council for Harmonisation 2016). Many countries reference this particular guideline in their national laws, which makes those guidelines legally binding. Hence, this set of conventions is not just one among the countless conventions in the field but is a particularly influential one. These conventions are, so to speak, the gold standard for what counts as acceptable data in clinical research. However, the pivotal epistemic role of these conventions for the reliability and trustworthiness of data has been largely overlooked by philosophers.

These conventions contain rules and principles that create accountability for the integrity of data and the safety of patients in a clinical trial. Their goal is to 'facilitate the mutual acceptance of clinical data by the regulatory authorities' (International Council for Harmonisation 2016, p. 1). The focus of the document is not the design of clinical studies; rather, it addresses the practical implementation of such studies. These guidelines define high-level principles about data quality and practical measures to implement them. That is, they define roles and responsibilities, mandatory instruments and documents, reporting duties and much more. As the document is structured according to roles and responsibilities rather than ideas, it presents a most unusual read for philosophers.

The high-level principle about data quality states that 'All clinical trial information should be recorded, handled, and stored in a way that allows its accurate reporting, interpretation and verification' (International Council for Harmonisation 2016, principle 2.10). Throughout the document, various responsibilities and measures are defined to implement this principle, encompassing data collection, documentation, verification and reporting.

The general epistemic approach to data quality entailed in this document is the approach of prospective quality *management,* meaning the prospective planning of data collection, controlling mechanisms and corrective feedback loops. The aim is to increase the likelihood of high quality data as defined in the document (see Figure 1). The first step is to plan and implement standardised data collection processes. To that end, investigators must have quality management systems in place, with well-established instruments. Examples are standard operating procedures that define the processes to generate, record and report data in line with the protocol and other regulatory requirements. Although this is a somewhat trivial point, the prospective and tailored control of data collection procedures in line with the specific purpose constitutes the essential epistemic difference between data that is generated according to the rules of GCP and data that is repurposed from routine healthcare. By contrast, retrospective quality assessment of repurposed data is mostly limited to describing the quality of existing data.

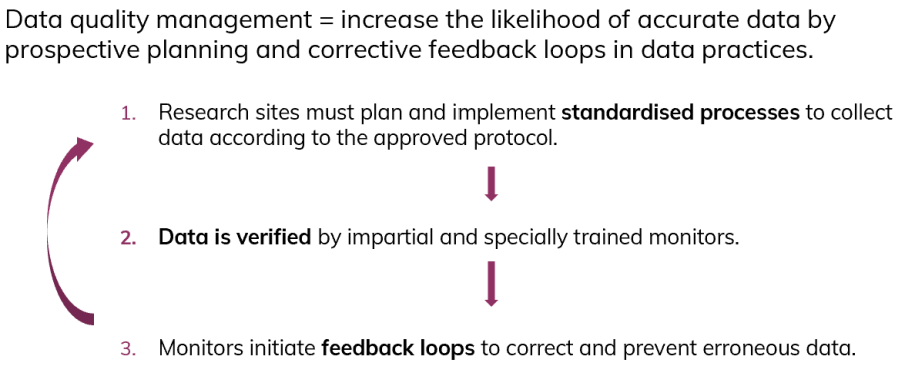

Figure 1: The process of data quality management in clinical trials.

The next step is more interesting and is unique to human research. Under the GCP guideline, data must be verified by independent and specially trained experts called 'monitors'. Following GCP, monitoring fulfils three purposes:

The purposes of trial monitoring are to verify that:

(a) The rights and wellbeing of human subjects are protected.

(b) The reported trial data are accurate, complete, and verifiable from source documents.

(c) The conduct of the trial is in compliance with the currently approved protocol/amendment(s), with GCP, and with the applicable regulatory requirement(s). (International Council for Harmonisation 2016, 5.18.1)

Item (b) explicitly charges monitors with the responsibility to ensure that data is accurate, complete and verifiable, i.e., to ensure that data is of high quality. In Chapter 6, I argue that the purpose stated under (c) makes an essential contribution to the trustworthiness of the data used in clinical trials. Here, I focus on the role of monitors for assuring data quality as stated in item (b). The extent of monitoring activities is negotiable and depends on the risks involved in a specific trial. However, the standard approach is for monitors to visit, in person, the sites where the research is conducted, at the beginning of, during and after the trial. On site, monitors have various responsibilities (the document lists 17). Among them – and in line with the overall purpose – they are charged with 'Checking the accuracy and completeness of the CRF [case report file] entries, source documents and other trial-related records against each other.' (International Council for Harmonisation 2016, 5.18.4 (m)) The case report form is a purposefully designed data collection form for each research participant. The guideline specifies what exactly monitors should verify:

(i) The data required by the protocol are reported accurately on the CRFs and are consistent with the source documents.
(ii) Any dose and/or therapy modifications are well documented for each of the trial subjects.
(iii) Adverse events, concomitant medications and intercurrent illnesses are reported in accordance with the protocol on the CRFs.
(iv) Visits that the subjects fail to make, tests that are not conducted, and examinations that are not performed are clearly reported as such on the CRFs.
(v) All withdrawals and dropouts of enrolled subjects from the trial are reported and explained on the CRFs. (International Council for Harmonisation 2016, 5.18.4 (m, i-v))

To check whether data collection adheres to these requirements, monitors verify data entries against source documents (such as laboratory test results), conduct consistency checks within the database or interview trial personnel. In case of any irregular findings, monitors are obliged to inform investigators about any data entry error, omission or illegibility and ensure that 'appropriate corrections are made, dated, explained and initialled' by an authorised person (International Council for

Harmonisation 2016, 5.18.4 (n)). The overall result of their activities is documented in a monitoring report with recommended actions for improvement. Thus, monitors not only detect issues in data quality but initiate step 3 of the quality management process. Feedback loops ensure that erroneous data is corrected and that standard operating procedures are adjusted to ensure that errors are not repeated. In case of repeated non-compliance or repeated lack of follow-up on a monitor's findings the research site is excluded from the trial. Quality management with on-site monitoring and the verification of source data poses a high bar for data accuracy and completeness: empirical evidence indicates that data that was completely verified by monitors on site can have error rates as low as 0.27%, with variability depending on the type of data and ranging from 0.0% to 0.36% (Andersen et al. 2015).

The ICH GCP Guidelines aim to foster mutual acceptance of data as evidence by researchers and regulators alike across countries. I have here provided a 'snapshot' of a 60-page document, but it is sufficient to illustrate that the bar for acceptable data in the field is set high. Not only is such data collected through tailored processes, it is also verified by specially trained experts and appropriately corrected by specially authorised personnel. While the prospective planning of data collection is a familiar practice in many fields, the role of specially trained monitors to verify data at the point of data capture is unique to human research. Due to such mandatory data collection and verification procedures, the data is relatively likely to be good in the sense that it is highly accurate and complete.

The application of these guidelines is in principle not limited to randomised trials; nor is their implementation methodologically necessary to conduct a randomised experiment. In previous chapters, I discussed the method of pragmatic trials. These are randomised trials that can measure outcomes through real-world data sources like electronic health records, the production of which was not governed by the rules of GCP. How to implement appropriate monitoring in such trials that ensures data quality while respecting the pragmatic attitude to leave the natural therapeutic situation intact is a matter of ongoing discussions (Simon et al. 2019; Irving et al. 2017). Hence, if one is looking for a criterion that sets real-world evidence apart from the gold standard as we know it today, it is not the familiar distinction between randomised and observational studies. Rather, it entails the processes that govern data collection and handling.

2. Bad data: Problems with real-world data

'Real-world data' is an umbrella term for all types of data generated *outside* a controlled experimental environment. Examples of sources include electronic health records, claims data, administrative data or registry data. The recording of health-related data outside experimental contexts is generally driven by the local needs, incentives and constraints of healthcare and not by the needs of researchers. In the words of pharmacoepidemiologist Sebastian Schneeweiss such data is 'filtered through the sociology of health care systems' (Schneeweiss 2016, p. 263). Yet, researchers have different needs for data compared with the needs of other practitioners: researchers often need more data, standardised data, fine-grained data, and particularly accurate, complete or representative data, to name a few. Hence, it is generally acknowledged by stakeholders in clinical research that real-world data is 'bad' data, in the sense that it often does not fulfil the needs of researchers and might compromise the integrity of a scientific inquiry. Schneeweiss and Avorn provide a useful overview of the various problems that can occur along the data collection processes (Schneeweiss and Avorn 2005).

Leonelli has carefully studied the decisive epistemic work that data curators contribute to making data journeys possible. Many of these practices increase the quality of data in an essential way. This includes the formatting, cleaning, standardisation and classification of data as well as the annotation of data with metadata. Metadata provide information about the data's context of generation and is essential to interpret the data and recontextualise it for a new use (Leonelli 2016, chapter 1). Depending on the level of curation that health data has been subjected to, its quality can differ greatly with regards to properties like structuring, standardisation or annotation with metadata. However, data curation has only limited influence on other quality dimensions such as the completeness or accuracy of the data. To illustrate the difference, I make a simple example of a weight measurement. Data curation ensures that the weight of a patient is recorded at the place, with the format and unit defined by a data model. Data curation transforms individual records to conform with the rules of the data model and might highlight or delete entries that show implausible values (such as negative values). If information about the context is available, data curators could add metadata about

how the measure was taken (with or without cloths)[16]. However, data curation has limited influence if a weight measurement was never recorded or deliberately exaggerated. I say a bit more what this limited influence is in the next chapter. Researchers and other stakeholders in the field are most concerned with those data quality issues that cannot be fully addressed by data curation. Such quality issues can generally not be corrected. Instead, the approach is to describe the extent of the problem and assess whether the use of the data for a specific purpose can tolerate the problem or not. This is what 'fitness-for-purpose' assessments are about. Here, I am taking the perspective of researchers and stakeholders and focus on the quality problems that cannot be fully addressed by data curation.

The list of potential epistemic problems that are inherent to such data is long (see, for example, Bower et al. 2017; Farmer et al. 2018). I briefly introduce three of the most common issues. The first problem is misrepresentation. If data collection is driven by the constraints and needs of insurers or healthcare providers, factors such as reimbursement incentives, time constraints or insurance coverage become part of the set of causal factors that are represented in the recorded data. Indeed, real-world data from electronic health records is often said to represent reimbursement realities rather than patient realities. An example is the 'upcoding' of diagnoses, which refers to physicians documenting a more complex diagnosis than the actual diagnosis for reimbursement purposes (Daniel et al. 2018; Schneeweiss 2019; Schneeweiss 2016).

Missing data values is another ubiquitous problem in routine data sources. As data values are collected on an 'as needed' basis, even key data elements are often missing for patients if they are not clinically useful. Tempini and Teira illustrate this point with the example of the Welsh Electronic Cohort for Children, where data about maternal smoking was found to be missing in 50% of the cases (Tempini and Teira 2020, p. 217). These concerns were also evident in a study that exposed potential data quality issues in the Scientific Registry of Transplant Recipients (SRTR). Yanik et al. evaluated the completeness of the SRTR data regarding the reporting of cancers in transplant recipients. Cancer reporting is mandatory in the SRTR database because transplant recipients have an increased risk of cancer due to the lifelong use of immunosuppressive therapy. However, the

[16] As Leonelli has shown these processes are far from trivial transformation processes but shape the evidential value of data and subsequent research greatly. This is particularly due to the choices for semantic standards, data models or metadata but also, as shown in her later work, by data security measures, see Tempini and Leonelli (2018).

researchers found that only 36% of cases reported in other cancer-specific registries were also reported in the SRTR database (Yanik et al. 2016).

A third common problem is the insufficient coverage of such data. While completeness pertains to missing values for individual patients, coverage pertains to absence of entire data fields or patient populations on a database level. Coverage has been empirically studied in meta-research that evaluated the feasibility of replicating clinical trials with real-world studies. One study examined 220 clinical trials published in 2017 and found that only for 15% of these trials were all key data elements available in routine data sources. The other 85% of the trials measured an intervention, an indication or an outcome that was unlikely to be recorded accurately in the routine data (Bartlett et al. 2019). Another study limited the feasibility check to post-approval trials. These are trials that the FDA requires because an investigational drug was licensed based on an accelerated approval programme. Such post-approval trials are likely candidates to be replaced by real-world studies under the FDA's Real-World Evidence Programme. However, the authors' conclusion is even more detrimental: for none of the 50 post-approval trials between 2009 and 2018 were all of the necessary data available in routine sources (Wallach et al. 2021).

A related concern is about the underrepresentation of healthy patients in routine data. In routine settings, data collection is initiated by patients' needs to contact healthcare services, which is often associated with the patient's health status. This leads to the problem that healthier patients, mild symptoms or the simple worsening of an existing problem are commonly not covered or at least underrepresented in routine data sources (US Food and Drug Administration 2021b).

Based on the various risks associated with routine data, many scholars warned against its use for research. Tempini and Teira contrast these risks with established principles for data quality and conclude that 'what counts as data depends on the risk threshold one works with' (Tempini and Teira 2020, p. 219). It is widely accepted that routine data cannot uphold the quality standards of good data according to the rules of GCP. However, a key critique of the GCP guidelines is that they overcontrol the quality of data by imposing a rule-following attitude that unnecessarily increases the costs of clinical trials (Collins et al. 2020). The FDA and other regulatory agencies are aware of the risks posed by real-world data while at the same time intrigued by the promises of real-world data. The solution to deal with data quality issues in real-world data is a contextualised approach to data quality. Data quality assessment frameworks help to systematise the process of

identifying data that is just 'good enough' for use in research – or 'fit-for-purpose'. Consequently, standardised frameworks to retrospectively assess the quality of data have become an indispensable piece of the puzzle for making repurposing efforts viable. Regulators and stakeholders in the field are currently developing such guidance and frameworks (European Medicines Agency 2022; US Food and Drug Administration 2021b). My aim in this Part II is not to show that real-word data is generally bad data. It is. My interest concerns data that has passed a data quality assessment test and is deemed fit-for-purpose for a specific research study. Throughout Part II of this thesis, I explore the epistemic risks involved in these assessments from different angles beginning by conceptual reflections on the notion of 'fitness-for-purpose' in section 3.

Before I turn towards this discussion some broad implications of the quality of routine data are worth noting. The current state of 'bad data' as largely unavailable and incomplete creates epistemic tensions and substantially limits the settings in which such studies are suitable. As a response to these limitations of routine data, we are likely to witness two transformations of research and healthcare. Firstly, routine data will affect what clinical studies look like and which research questions are asked. Studies will move towards more hedged clinical research questions, broader patient populations and more robust outcomes because routine data would otherwise be unavailable. Whether such conceptual transformations are acceptable to those who rely on such evidence will require substantial negotiation on a general and contextual level. The shift from surrogate outcomes towards clinically relevant and robust outcomes, or the shift towards broader patient populations, will be welcomed even by critics. Other conceptual shifts may be more controversial. I discuss the conceptual shift in the notion of a medical intervention that accompanies pragmatic trials in Chapter 3. Here, I want to note that such conceptual shifts will be fairly common in real-world studies for all sorts of concepts involved, and I provide further examples in the case study below. Secondly, aspirations to use routine data for various purposes will change data collection and data handling at the source – in the clinics. Because the needs and incentives for data collection in routine settings diverges from data needs of science, some countries have already implemented new policies or programmes to incentivise the collection of data to make data reuse possible. Green and colleagues trace the practical ethical consequences of such 'data work' for health care in Denmark (Green et al. 2022), and I discuss their contribution in Chapter 5. This second transformation clearly belies the idea that routine data is 'secondary use data' that researchers can simply reuse at minimal additional costs. Moreover,

routine data might have the advantage of being representative of a wide variety of patients, yet this is often no longer the case once the data has been cleaned of all potential quality issues. Worse, the subset of patients with sufficiently high-quality data could represent generally less-healthy and better-cared-for patients. This concern has been raised repeatedly (Schneeweiss 2016) . With the current quality of real-world data, a phenomenon like the trade-off between internal and external validity occurs: Data that is of sufficient quality to generate unbiased evidence may represent a highly selective subset of patients. This tension belies what is believed to be one of the main epistemic strengths of such data. The generally 'bad' quality of real-world data is not a general obstacle for their use. However, the current 'bad' quality of such data plausibly decreases two of its main purported advantages.

3. Good enough data: Data quality as fitness-for-purpose

Traditional data handling practices were built on the assumption that data and its purpose are inextricably intertwined. For example, data protection laws often only allow the collection of data for a specific purpose. GCP guidelines therefore do not need to reflect about the purpose of data, because the existence of the data itself presupposes its purpose. In the realm of big data and data repurposing *fitness-for-purpose* is a widely adopted success criterion for data quality across scientific disciplines (Floridi and Illari 2014). With the advent of routine data, clinical research recently also turned towards such an approach. For example, the EMA adopted the criterion in their data quality framework (European Medicines Agency 2022), the FDA sometimes uses the related notion of 'fitness-for-use' in their guidelines (US Food and Drug Administration 2021b) and the revised ICH Guidelines on General Considerations for Clinical Studies adopted the notion as an overall quality criterion for clinical studies (International Council for Harmonisation 2021b). Such data quality frameworks rarely reflect on the meaning of fitness-for-purpose but proceed to disentangle the notion into 'quality dimensions' and propose metrics to measure them. However, this has been proven a notoriously complex tasks with a wide lack of convergence on how to best approach it (Illari 2014). The first systematic data quality framework for EHR data was developed only in 2013 by Weiskopf and Weng. A widely used framework was published Kahn and colleagues in 2016 (Kahn et al. 2016; Weiskopf and Weng 2013).

Clinical research has a well-established tradition of quality assessment for the design of clinical trials, examining properties such as randomisation or drop-out rates. The question of data quality assessment is, however, relatively new. Within the philosophy of biomedicine, engagement with the notion of data quality or practices to ensure data quality is likewise relatively scarce. Stegenga approaches the question of information quality from the perspective of risk-of-bias assessment tools for clinical trials (Stegenga 2014). Canali turns to examples from biomedicine to argue for a general turn towards contextual data quality assessments (Canali 2020). Leonelli discusses six methods of data quality assessment in biology: a review of the data by peers; the involvement of data curators and investigators; review of data through automated processes; crowdsourced assessments; ratings by future data users; and the reliance on specific technologies for data production. She explicitly argues against tying data quality to the use of particular technologies to avoid that quality data require a resource rich environment (Leonelli 2017b). Philosophers of information and computer sciences have critically engaged with detailed epistemic questions of data quality frameworks such as those mentioned above (Floridi and Illari 2014). Bokulich and Parker recently developed an adequacy for-purpose account for data models (Bokulich and Parker 2021).

In this section, I reflect on conceptual issues of the notion of fitness-for-purpose and illustrate the difficulties to bridge the gap between this abstract notion and concrete criteria to measure it. Defining the notion's success criteria involves various normative and theoretical commitments about the purpose of a study, the necessary conditions to achieve the purpose and the uncertainty that can be tolerated in this endeavour. Some interpretations of the notion might almost imply a free pass for the use of routine data while others might set the bar so high that it is nearly impossible for routine data to meet the requirements. Chapter 5 focuses on methodological issues of how to measure and evaluate data quality.

Philosophers have thought about adequacy-for-purpose in the scope of scientific models for a while. Bokulich and Parker recently explicitly proposed adopting an adequacy-for-purpose criterion for data models (Bokulich and Parker 2021). In the adequacy-for-purpose view on scientific models, the success of models is not solely determined by how closely a model represent their target, but rather whether the model can be used to achieve a desired purpose. The same criterion can be applied to data or data models. Data quality then is a part of a 'larger problem space' where evaluating data models becomes a question about how well data 'stands in a suitable relationship with the representational target, a data user, and available methodologies' (Bokulich and Parker 2021, p. 11).

Following Bokulich and Parker, we can define adequacy-for-purpose as follows: '[A] dataset or data model D is adequate for purpose just in case the use of D in instance I would (or would be very likely to) result in the achievement of P' (Bokulich and Parker 2021, p. 11) . Fit-for-purpose, in their account, is simply the graded notion of adequacy-for-purpose. The concept can be used in cases where a certain purpose can be achieved to a greater or lesser extent. Bokulich and Parker also propose distinguishing between two meanings of 'adequacy-for-purpose', namely, adequate-in-an-instant and adequate-given-resources. The second term refers to situations where the adequacy might depend on access to an extended set of resources, such as technologies to successfully use the data and achieve the intended purpose. For example, data might not be adequate-in-an-instant if the researcher lacks access to a reference dataset that would help to correct errors in the data. However, the same data might be adequate-given-resources if one has access to such reference data. The distinction is not always clear-cut, because using data successfully always requires certain resources. The distinction however hints at the interesting aspect that there is a wide range of additional resources that could be used together with a set of data. As I show in my case study in the next section, the FDA inexplicitly included various additional resources in their assessment to make up for potential shortcomings of the data.

To put the criterion to use, researchers require first a good understanding of the purpose and the epistemic and non-epistemic properties of data needed to achieve that purpose. The common purpose of data is to answer a research question. Yet, as Bokulich and Parker argue, the ultimate purpose could also be a practical aim that is only mediated by the epistemic aim of answering a question (Bokulich and Parker 2021). In the context of clinical trials, the purpose of data is often to produce evidence to enable regulatory decision-making or healthcare decision-making. The ICH recently revised their Guideline on General Considerations for Clinical Trials and explicitly defined the quality of a clinical study as fitness-for-purpose. They briefly reflect on the purpose of a study as follows:

> The purpose of a clinical study is to generate reliable information to answer the research questions and support decision-making while protecting study participants. The quality of the information generated should therefore be sufficient to support good decision-making. (International Council for Harmonisation 2021b, p. 6)

Following the ICH's reflection, data quality depends on requirements of data for good (regulatory) decision-making. Depending on one's view what good

decision-making entails, data might require different epistemic and non-epistemic properties to be fit-for-purpose. The traditional GCP approach is to rely on data from highly controlled trials with highly accurate and complete data. Pragmatists, however, might hold that highly accurate data from ideal clinical trials is epistemically inadequate to support good decisions, because these data are not the right kind of evidence to support the use of treatments outside the context of the trial (for a discussion see Chapter 3). Irving et al. for example have argued that the rules of GCP impact how well data represent routine clinical care which compromises generalisability (Irving et al. 2017). The purpose dependency of data quality runs deep. Illari argues that purpose-dependency of data quality can even affect the meaning of quality dimensions, such as accuracy – and I agree (Illari 2014). The concern for pragmatism illustrates her point. If measuring 'real-world' effects is the purpose of a study, 'bad' data could be accurate data precisely because it represents the local variability one is interested in. By contrast, 'good data' would be inaccurate in this context because it does not accurately represent local realities.

Furthermore, Bokulich and Parker propose that the adequacy-for-purpose of data models can also depend on pragmatic criteria analogous to the adequacy-for-purpose of scientific models. In their view, pragmatic criteria could incorporate the need for fast or cheap availability of data into fitness-for-purpose evaluations (Bokulich and Parker 2021). The FDA's criteria of provenance – the property that data can be traced back to its origins – can be understood as such a pragmatic criterion.

For putting the fitness-for-purpose criterion to use, researchers also require a good understanding of both the degree of informativeness and the acceptable uncertainty to achieve the purpose. Unlike data controlled by GCP, routine data is generally not fully accurate and complete. The fitness-for-purpose approach acknowledges that perfect accuracy and completeness might not be required to achieve a certain purpose. Instead, all that is needed is that data are *sufficiently* accurate, complete, relevant and others. Consequently, researchers need to determine thresholds for sufficiently accurate, complete or relevant data for achieving the intended purpose. There are several ways to approach this. A way to define such a threshold recommended in the FDA's guidelines is by saying that a level of uncertainty is acceptable as long as it does not change the interpretation of the results. Another way to approach the issue is by balancing uncertainty in light of the social and ethical implications of error. Thus, data might be said to be sufficiently complete if it is meant to answer a question with no ethical stakes attached. A third approach to accommodate inaccurate data is by hedging the

scientific hypothesis. To use the example of Bokulich and Parker: A rain gauge reading that shows 40 mm on average, but is known to systematically overestimate the amount of rain, can still be used to reliably answer the question of whether the rain was less than 100 mm, while it cannot be used to estimate the exact amount of rain (Bokulich and Parker 2021). Here, inaccurate data is sufficiently accurate if the question asked is sufficiently hedged. Rephrasing effectiveness claims in terms of therapeutic actions rather than pharmaceutical properties of drugs, as I proposed for pragmatic clinical trials in Chapter 2, is another example of hedging to account for uncertainties in the data.

Operationalising fitness-for-purpose is a complex task. The reflections on the notion show that its success criteria depend on various normative and theoretical commitments. They include one's view what good decision-making entails, the epistemic and non-epistemic properties of data to achieve a specific purpose as well as commitments about the necessary level of informativeness and certainty. Nonetheless, Illari holds that fitness-for-purpose judgements are deeply relational but not subjective judgements. That is, once the user has determined the purpose, evaluating whether some dataset is fit-for-purpose only depends on objective judgement. Even if this view is correct under some sense of objectivity, without making explicit what the entailed commitments are, the notion does not sufficiently constrain what data might pass the fitness-for-purpose test. Some of these commitments almost imply a free pass for the use of routine data while others might set the bar so high that it is nearly impossible for routine data to meet the requirements.

Data quality frameworks usually disentangle the notion into various *quality dimensions* (such as accuracy and completeness) as well as data quality categories (such as intrinsic, foundational and contextual). The main issue that philosophers and scientists have identified with these attempts is that there is little convergence on how to divide the concept of data quality into quality dimensions and quality categories (Illari 2014; Weiskopf and Weng 2013). To illustrate, the list below shows the quality dimensions and their definitions suggested in a pioneering framework by Weiskopf and Weng.

- **Completeness:** Is a truth about a patient present in the EHR?
- **Correctness (accuracy, reliability):** Is an element that is present in the EHR true?
- **Concordance (coherence):** Is there agreement between elements in the EHR, or between the EHR and another data source?
- **Plausibility:** Does an element in the EHR make sense in light of other knowledge about what that element is measuring?

— **Currency (timeliness):** Is an element in the EHR a relevant representation of the patient state at a given point in time? (Weiskopf and Weng 2013, p. 145)

A recent review of data quality frameworks found that more recently, researchers have expanded the frameworks to include the following domains and definitions (Bernal-Delgado et al. 2022):

— **Relevance**: pertaining to the availability of data items in a data model
— **Coverage**: pertaining to the population and timeframe that is covered by a database

Finally, the FDA added domains that are particularly relevant in the regulatory context, along with the following definitions in each instance:

— **Provenance:** An audit trail that accounts for the origin of a piece of data (in a database, document or repository) together with an explanation of how and why it got to the present place.
— **Traceability:** Permits an understanding of the relationships between the analysis results (tables, listings, and figures in the study report), analysis datasets, tabulation datasets, and source data.

The lack of convergence in the frameworks for data quality is an obstacle for scientists to make progress in the field because it hampers the sharing of experience and expertise (Illari 2014). Leonelli has likewise noted that the diversity of data quality assessments make it difficult to establish international standards (Leonelli 2017b). Some scholars have argued in favour of abandoning the search for data quality dimensions entirely, because the dimensions do not perform any useful function for the development of concrete metrics to measure the data quality (Embury and Missier 2014). Others attempt to support researchers with a new approach to mapping quality dimensions that overcomes the conceptual confusions that have hampered progress (Illari 2014). Thus, although there exists an abundance of such frameworks, they are barely helpful to guide scientific practice.

4. Case study: FDA approval of Prograf

In this section, I introduce the second case study to illustrate the epistemic opportunities and risks that the contextualised approach to data quality entails. It is the first study submitted to the FDA under their new real-world evidence programme. The case is fascinating because the kind of evidence submitted to the FDA diverges considerably from the gold standard requirements. At the same time,

the study is neither particularly risky nor innovative. It appears that the researchers achieved their purpose with considerable efficiency – and probably a piece of luck. The case illustrates the opportunity this approach entails to recognise various forms of evidence as sufficient for good decision-making. However, it also raises various serious concerns about the risks entailed for future applications.

Prograf is the proprietary name of an immunosuppressant drug widely used to prevent rejection of solid organ transplantation. It contains the active ingredient tacrolimus, a calcineurin inhibitor (CNI) that inhibits T-lymphocyte activation. Despite there being a variety of immunosuppressant available, CNIs are a main component of most immunosuppressive regimens. Tacrolimus was initially approved in the US for liver transplants in 1994 and is marketed by Astellas Pharma US, Inc. In July 2021, the FDA decided to extend the label to adult and paediatric patients who receive lung transplantation – a paradigm decision which was exclusively based on real-world evidence. On 16 July, the FDA announced in a press release:

> Today, the U.S. Food and Drug Administration approved a new use for Prograf (tacrolimus) based on a non-interventional (observational) study providing real-world evidence (RWE) of effectiveness. ... This approval reflects how a *well-designed*, non-interventional study relying on *fit-for-purpose* real-world data (RWD), when compared with a *suitable control*, can be considered *adequate and well-controlled* under FDA regulations. (US Food and Drug Administration 2021a, my emphasis)

An amendment to the Food, Drug and Cosmetics Act in 1962 introduced the requirement that only 'adequate and well-controlled clinical studies' were acceptable as evidence for regulatory approvals. This amendment marked the advent of the RCT becoming the regulatory gold standard for such assessments. Later, a more specific interpretation of the requirement was included in the US Code (21 CFR 314.126) (Teira 2020). The amendment to the US Code on the use of real-world evidence, which came with the 21st Century Cures Act, states that the use of real-world evidence should not be interpreted as undermining these well-established quality criteria. The FDA's press release on 16 July clearly stated that the approval of Prograf met these evidential requirements because the submitted study was 'well-designed', was based on 'fit-for-purpose data' and used a 'suitable control'. In their publication, the research team mentions the use of real-world evidence even as one of two main strengths of the study (Erdman et al. 2022, p. 1241). I examine the available evidence behind these claims and identify relevant epistemic concerns. I argue that accepting such evidence as proof of

efficacy cannot be explained by the quality of the study and the submitted data alone; instead, additional contextual evidence did most of the epistemic work.

Before delving into the case, I provide a few more helpful details about the regulatory context. At the centre of attention in the FDA's Real-World Evidence Programme is the use of such evidence to support the efficacy of treatments for new indications and to fulfil post-approval requirements.[17] The approval of a new indication means that the use of medicines that are already available on the market, after a regular approval procedure, is extended for treating another disease or for another population or for prescribing a different dosage. A post-approval requirement means that the medicine is granted approval that is conditional on additional evidence about safety and efficacy. The most innovative component of the current development concerns the *exclusive* use of real-world evidence for efficacy estimates, i.e., its use to *substitute* clinical trials and not merely to supplement trials. The use of such evidence in a supplementary way – to partially replace data collection within clinical trials or in post-marketing safety studies – has a longer tradition. Studies about these experiences with real-world data for these supplemental uses have been published (Franklin et al. 2020b; Jonker et al. 2022; Mahendraratnam et al. 2022). Hence, when the FDA announced its first approval of Prograf based on real-world evidence, this was an approval to *extend* the indication of tacrolimus for use in lung transplantations, *exclusively* based on real-world evidence.

My analysis of the case is based on the following available materials. First, the FDA assessed the effectiveness of tacrolimus through a multi-discipline review. The results were published together with additional reviews and the label (rules on how to prescribe and administer the drug) on the FDA's website (CDER 2021). The FDA also published a press release, and expert discussions were held in which

[17] The relevant section of the 21st Centuries Cures Act (21 CCA) is section 3022, stating 'The FDA must evaluate and issue guidance on the use of evidence from sources other than clinical trials to support approval of a drug for a new indication.' Section 3022 of the 21 CCA initiated the development of section 505f of the Federal Food, Drug and Cosmetics Act titled Utilizing Real-World Evidence (US Code, Title 21, Chapter 9, Subchapter 5, §355g), which defines 'real-world evidence' and further develops the mandate to say that the two purposes for which the use of real-world evidence must be evaluated are: 'a) to approve of a new indication of an already approved drug' and 'b) to fulfil post approval study requirements'. Available online: https://www.congress.gov/bill/114th-congress/ house-bill/34?s=2&r=34 last accessed 10 October 2021. The article by Fraile Navarro et al. (2021) on the Salford Lung Study was so timely that they refer to an earlier version of the 21CCA (H.R.6 instead of H.R.34), where the relevant section 2062 still said '[the FDA] must evaluate the use of evidence from clinical experience (in place of evidence from clinical trials) and establish a streamlined data review program.' The explicit mention of 'clinical experience' is an unfortunate wording that did not pass into law.

stakeholders and third parties discussed details of the case (US Food and Drug Administration 2021a). The research team published their findings in the journal *Transplantation* (Erdman et al. 2022). Second, my analysis draws from a webinar that was hosted by the real-world evidence special interest group of the Statistics in the Pharmaceutical Industry Association (PSI RWD SIG 2021). In the webinar, various stakeholders involved in the planning, conduct and review process of the case discussed details about the process. The speakers were:

- Richard Croy, Global Statistical Lead at Astellas Development
- David Nimke, Real-World Evidence US Lead at Astellas
- Josi Wolfram, Clinical Development Applications Lead at Astellas
- Tim Weaver, Statistician at Chronic Disease Research Group (CDRG), operating the SRTR
- Tae Hyun (Ryan) Jung, Biostatistics Reviewer at the FDA

When quoting from the webinar, I refer to the speaker by their last name and indicate the time-point in the video in brackets. No written transcripts or audio-visual recording are available from the FDA Advisory Committee because the extension approval of Prograf did not require the involvement of the Committee, whose role is to advice only on controversial cases.

4.1. The primary evidence submitted to the FDA

'A well-designed, non-interventional study'

The approval of the medicine was based on an observational study that the FDA evaluated as being 'well-designed'. The study included 20,080 US patients, during nearly a 20-year period (1999–2017), who received lung transplantations. The primary outcome was a composite outcome of either death or graft failure after one year; it allowed for estimating a clinically highly relevant outcome of overall graft survival after one year. Because patients usually receive tacrolimus as a regimen, which means a combination of different immunosuppressive drugs, the effectiveness of tacrolimus was measured as a regimen rather than an individual component. The study had few inclusion and exclusion criteria, and only 12% of patients were excluded from the study because of the criteria.

A main complexity of the design and statistical analysis was a problem that is common to observational studies, called 'immortal time bias'. This bias usually happens when researchers assign participants to treatment groups based on an observation from after the beginning of the study, such as the filling of a

prescription. In this case, the beginning of the study was intended to be on the day of the transplant; however, data on the patient's treatment regimen was available only at the day of discharge from the hospital. Patients must have survived the time lapse between study entry and group assignment in order to be allocated to the treatment group. Hence, they contributed 'immortal' time to the treatment group. Once recognised, immortal time bias is usually simple to correct for. The Astellas research team had two options to account for the problem. They chose one option as their primary analysis (left truncation), which constituted the primary evidence; the other option was used as a sensitivity analysis (incident user design). Sensitivity analysis is used to test the robustness of results, whereas the primary analysis provides the main results for the assessment.[18] The team's reasoning for choosing one option over the other was interesting:

> The reason for choosing left truncation over the incident user design was based on the idea that we would be able to have language in our label in the future that would be consistent with the way clinical trials were designed so we would be able to say the risk of some outcome post-transplant was this, that's a little bit different than saying the risk of some outcome post-discharge date is this other estimate. (Nimke, 34:35)

Hence, the team wanted to have effectiveness claims from their study that are conceptually familiar from clinical trials. In a prospectively planned clinical trial researchers would generally estimate the treatment effect of the immunosuppressant drug directly after the transplantation. One of the two solutions to analyse the routine data could only estimate the treatment effect after hospital discharge, therefore they opted for the other solution. Unfortunately for the team, the solution they chose for their primary analysis resulted in an uninterpretable and imprecise estimate; hence, the team conducted another post-hoc analysis to account for the difficulties. Ultimately, by decision of the FDA, the results of the sensitivity analysis (which represented the second solution to the immortal time bias and were consistent with the results of the post-hoc analysis) were accepted as the primary analysis and reported in the label.

[18] A sensitivity analysis is conducted to test the impact of key assumptions in the statistical design. If results of the sensitivity analysis are consistent with the primary analysis and support similar conclusions, a result is robust under different key assumptions. The primary analysis, however, is usually the one that constitutes the primary evidence about efficacy.

'Suitable control'

The team initially planned a comparative study to compare a tacrolimus-containing regimen with another regimen used in clinical practice. They suggested a comparative analysis using Cox modelling, which is commonly used to investigate associations between survival time and predictive risk factors. Surprisingly, in pre-approval discussions with the FDA, the FDA clarified that their decision would not be based on this comparative analysis. According to a speaker in the PSI webinar, the FDA indicated instead that 'the rate of death or loss of graft in the tacrolimus IR + MMF arm will be our primary consideration to determine the efficacy of this product' (Nimke, 36:13). Thus, the primary comparison considered by the FDA was not the comparative analysis but 'the well-documented natural history of a transplanted drug with no or minimal immunosuppressive therapy' (US Food and Drug Administration 2021a).

Wolfram elaborates that the FDA approached the team towards the end of the review period to ask for a 'literature based review of historical patient outcomes in the absence of immunosuppressive therapy, ... in particular to elucidate the contribution of tacrolimus to the triple regiment' (Wolfram, 49:15). Even more surprisingly, the main reason indicated by the FDA for the recommendation was that 'the non-randomised comparator groups might be 'too different' to be compared, even after adjusting for confounders' (Nimke, slide 30). The historical literature review, as reported in the multi-discipline review by the FDA, tells the story how – and partially why – transplants (including lung transplant) were made possible by the invention of the first CNI medicine, cyclosporine. It traces the history of lung transplantation in humans, based on the earliest case reports as well as nonclinical studies on animals, mechanistic evidence and comparative reasoning. The conclusion was that 'Both the nonclinical and the clinical lung transplantation outcomes using cyclosporine, clearly prove the contribution of cyclosporine to success which otherwise would not be possible' (CDER 2021, p. 14).

As mentioned above, the US Code not only states that studies must be adequate and well-controlled but also provides a more precise interpretation of this phrase under different circumstances. While the use of a historical control can be adequate under particular circumstances, the use of a historical literature review – including case studies on different treatments and indications – is quite unusual. Acknowledging this, the FDA reasoned that

> While historically controlled studies usually involve a treatment group with drug assignment according to a protocol (as in a single-arm clinical trial), the regulations do not require such a design when comparing outcomes of treatment and historical control groups. (CDER 2021, p. 16)

It is certainly puzzling that the FDA considered a historical literature review a 'suitable control' and sufficient for the study to count as 'well-controlled'. Even more puzzling is that such a control was deemed more acceptable than a comparative design that would adjust for confounders, because the groups could be 'too different' despite the adjustment.[19]

'Fit-for-purpose real-world data'

Many discussions about the quality of real-world evidence revolve around the quality of data. When communicating their decision, the FDA and other stakeholders repeatedly adopted the term 'fitness-for-purpose'. In one section in their press release, the FDA specifies that the notion of fit-for-purpose refers to the dimensions 'relevance' and 'reliability' of data, which are developed in detail in a guidance document on assessing the quality of real-world data (US Food and Drug Administration 2021b). I discuss the specifics of this document in Chapter 5 together with an in-depth analysis of the methods and reasoning used by Astellas and the FDA to evaluate the quality of the data. For now, I briefly introduce the database and expose the overt limitations of the data that the team was willing to accept.

The database used for the study was the US Scientific Registry for Transplant Recipients, (SRTR). As its name says, this is a registry database. Such databases collect observational data on a specific area, for example a disease or a medical device, and they are built to serve scientific and sometimes policy purposes. However, registries are intended to be agnostic towards specific research questions; they do not follow the strict rules for data collection and management that guide clinical trials. Therefore, they pose similar problems in data quality as other sources of repurposed health data, even though they often contain more disease-relevant information than other sources.

---

[19] The publication from the study did report the comparative treatment effects in terms of a proportional hazard ratio, one estimate unadjusted and two adjusted. The reported adjusted covariates are mentioned in a footnote alongside the results table and include a multivariable proportional hazard model adjusting for age at transplant, recipient sex, lung transplant procedure, transplant time period, diagnosis, BMI, race, ethnicity, LAS at transplant, serum creatinine (mg/dL) at transplant, eGFR [estimated glomerular filtration rate] at transplant, total bilirubin (mg/dL) at transplant, length of hospital stay (days), donor age group, donor race, lung total ischemia time (hours), donor–recipient weight ratio, donor–recipient CMV matching and induction with IL-2 receptor antagonists. Erdman et al. (2022, p. 1239) While the team acknowledges the possibility of residual confounding for unadjusted covariates, there is no discussion about the choice of covariates.

Since 1987, the SRTR has collected data on all solid organ transplants in the US, by law. The database is generally acknowledged as highly representative of the transplant population. It has been used by the Organ Procurement and Transplantation Network (OPTN) and the transplant community for decades, mainly to facilitate the regulation of organ donation and allocation in the US. It allows practitioners to register candidates for transplants, match donated organs to candidates and submit data on donors, candidates and recipients, both before and after the transplants. Moreover, the law requires the mandatory and comprehensive reporting of transplant procedures and outcomes into the registry, with a focus on the 'breath of data rather than its depth' (Waves, ca. 51:00).

The largest part of the data is collected by the OPTN and comes from various sources such as transplant programmes or histocompatibility laboratories. Such data is collected through specific forms and guided by OPTN policies on the timely collection of data. The data is supplemented by additional data sources, particularly the Death Master File at the National Technical Information Services. Both the robust operational structure and the use of a trusted external data sources were recognised by the FDA as contributing to the quality of data contained in the database. To assess the relevance of the data sources, the research team compiled a list of the data elements needed in an ideal scenario and assessed the availability of the data element in different data sources. They found that while none of the data sources contained all the data elements for the ideal case scenario, the SRTR was a 'research-driven data collection that is generally relevant to the disease, potentially representative although continuity might be lacking' (Nimke ca. 21:00).

In addition to the robust operational structure of the database, Astella collaborated with local experts regarding the database. The Chronic Disease Research Group (CDRG) at the academic Hennepin Healthcare Research Institute in Minneapolis has operated the SRTR database for decades. Astellas was using data over a 20-year period, during which data elements had changed. Faced with the decision to either insource or outsource the data management and analysis, they opted for outsourcing to the CDRG. The CDRG is responsible for providing statistical and other support to the OPTN, and in that role, CDRG works with the data to publish various publicly available reports. Examples are programme-specific reports and organ-specific reports. Acknowledging the added value of the GDRG expertise, the Clinical Development Applications Lead at Astellas confirmed that 'We figured having their experience and expertise with analysing these data would be a value add and I can say it was. ... they know the data collection and reconciliation steps best and so could better communicate considerations in

that regard' (Wolfram, 51:33). Leonelli and others who applied her framework have repeatedly highlighted the local expertise that is required to interpret and analyse repurposed data (Leonelli 2016). In this case, the investigators acknowledged this concern and planned their collaboration accordingly.

Despite the overall promising outlook of the database as potential evidence for the approval of Prograf, the researchers had to deal with considerable limitations. First, as discussed, the team had to account for the problem of treatment exposure data only being available from the day of hospital discharge rather than from the day of the transplant. The lack of this data traces back to the policies on the timely collection of data that govern the registry (Organ Procurement and Transplantation Network 2023). According to these policies, the data on which drug regimen patients received was collected on the Transplant Recipient Registration Form, which the data collection policy states must only be filled in at hospital discharge.

Second, for the primary outcome, data was available and was reported with exact dates, which allowed for the desired time-to-event analysis. However, the analysis of secondary outcomes was limited to coarse-grained time intervals because the data did not reflect the precise dates of these events. The reason traces back to the precise wording of the data collection forms and rules. For example, to assess whether a patient has been hospitalised, the follow-up form asks 'Has the patient been hospitalised since the last patient status update?' – which can be answered by either yes or no. Yet the time interval since the last patient status update spans usually an entire year (Waver, 50:00).

A third limitation of these data is that tacrolimus is always administered together with other immunosuppressants. Consequently, the data did not allow for estimating the treatment effect of tacrolimus in isolation but only as one component among others in a regimen-based therapy.

The final and most serious limitation concerned the dosage. Tacrolimus has a narrow therapeutic window, which means that the range of drug concentration in the blood between the minimal effective and the minimal toxic concentration is narrow. The dosing recommendation on the label therefore only suggests an initial dose together with recommended ranges to target trough levels that need to be tailored to patients (target tough levels are the desired minimum concentration of the drug in a patient's blood at the end of the dosing interval, just before the next dose is given). The problem is that 'large amounts of data are needed to support as such a label recommendation' (Croy 09:40). Obtaining such data was a challenge

for the submission because none of the databases the team could access, including the SRTR, contained data on target trough levels (Nimke 26:54).

Despite these shortcomings the FDA and Astellas considered the data fit-for-purpose. However, an undetected data-quality issue was probably the main reason for the problem with the primary analysis that resulted in an uninterpretable and imprecise estimate. Prompted by this unusual finding, the researchers scrutinised the data. They found that a few patients were reported for hospital discharge at the day of the transplant, which clinical experts think is implausible. Unfortunately, the method of analysis they chose depended on the number of patients who were discharged at the time of the first event and was highly sensitive to early events. During their review, the FDA was interested to understand the reasons for the data indicating these early hospital discharges. The team reasoned that it could either be an error in the data (e.g., a mix-up of dates) or could mean there was no proper distinction between patients who were discharged and those who were transferred to another hospital. All parties acknowledged that they could not go beyond speculation. Given the retrospective nature of the study; going back to the research site to verify what had happened was not an option. Thus, despite the vast data, a few patients with incorrect data about their hospital discharges were sufficient to render the primary analysis uninterpretable (CDER 2021, see section 8.3 Statistical Issues).

To make up for this problem, the team conducted a post-hoc analysis in which they set the dates of these early events to a later date. Both the post-hoc analysis and the sensitivity analysis that estimated the treatment effect after the day of discharge now provided precise treatment effects. The FDA decided to accept the sensitivity analysis as the primary analysis and granted approval for the medicine. The label now states

> 14.4 Lung Transplantation
> The efficacy and safety of PROGRAF-based immunosuppression in primary lung transplantation were assessed in a noninterventional (observational) study using data from the US Scientific Registry of Transplant Recipients (SRTR). The study analyzed outcomes based on discharge immunosuppression treatment regimen in recipients of a primary lung transplant between 1999 and 2017 who were alive at the time of discharge. In adult patients receiving tacrolimus immediate-release products in combination with MMF (n=15,478) or tacrolimus immediate-release products in combination with AZA (n=4,263), the one-year graft survival estimates from time of discharge were 90.9% and 90.8%, respectively. [...] (CDER 2021, 14.4)

Let's take stock. The SRTR is generally a highly relevant, disease-specific and well-established database. It contains data on all transplant patients in the US since 1987, among whom almost all patients during a 20-year period were included in the study. The primary outcome was a robust outcome and was reliably captured in the database. However, the team accepted the data as 'fit-for-purpose' despite the known and relevant limitation that dosing recommendations were unavailable in the data and treatment effects could only be estimated for entire regimens, and only at the day of hospital discharge. Moreover, an important data-quality issue went undetected. To account for the failed analysis, the team made a post-hoc switch between the primary and sensitivity analyses.

Confidence in accepting the evidence as 'adequate and well-controlled' proof of efficacy clearly cannot be explained by the design being 'well-controlled', a 'suitable control' or the 'fit-for-purpose real-world data' alone. Indeed, in their multi-discipline review, the FDA review team provided further justification for accepting the study as fulfilling the 'adequate and well-controlled' requirements by referring to additional contextual evidence. I identify three kinds of contextual evidence that further supported the evidence and the FDA's assessment that the evidence was 'adequate and well-controlled'. Such evidence was the unambiguous natural history of the disease, the large effects and the sufficient comparability of other indications.

4.2. Three kinds of contextual evidence

Unambiguous natural history of disease

In the multi-discipline review, the FDA justified the use of the natural history as the control in reference to the requirements for well-controlled studies as outlined in the US Code. These requirements mention two examples where historical controls can be considered adequate: diseases with 'high and predictable mortality' and treatments in which 'the effect of the drug is self-evident'. A brief look at the history of transplants shows that lung transplantation is a clear case of the first kind. For example, the historical literature review mentions a non-interventional study with 36 patients in the late 1960, with a median survival of less than two weeks and no patients surviving to one year. Wolfram commented on the results from their historical literature review: 'Essentially with no or with minimal immunosuppressive therapy, the graft survival rate is – well there isn't really a survival rate' (Wolfram 49:52). Not only was mortality high but the outcome also occurred rapidly, within a few weeks. With such high and rapid mortality in the absence of treatment, there is no need to account for the natural remission of

transplant patients in the effect estimate. This point limits an important source of bias in the study and the effect estimate.

Large effects

The FDA reported in their press release that in the study 'a dramatic improvement in outcomes was observed among lung transplant patients receiving Prograf as part of their immunosuppression medications'. They did not understate that point. The graft survival rate for the two tacrolimus-containing regiments after one year were above 90% in adults and slightly lower for paediatric patients. Mortality without immunosuppressive medicines is around 100%, and the effectiveness of the treatment regimens is thus remarkable. Experts widely acknowledge that the impact of various biases is irrelevant given such large effect sizes, because it is implausible that mere bias could explain such strong treatment effects. Therefore, the absence of bias in such cases would not change the interpretation of the results. Accordingly, the FDA wrote as follows in their multi-discipline review:

> The division acknowledges that given the lack of contemporaneous data collection, differences may exist between the treatment and the historical control groups. The SRTR data go back to 1999 and some differences would be expected between patients in the SRTR and historical controls due to changes in baseline characteristics of patients receiving transplantation, surgical techniques, and supportive medical care over time. Nonetheless, the clinical benefit seen with the tacrolimus-containing immunosuppressive regimen studied is so large compared to historical controls that differences in baseline characteristics, surgical technique, and/or supportive care between groups are highly unlikely to explain the outcome differences, and therefore do not change the conclusion of effectiveness. (CDER 2021, pp. 17–18)

Sufficient comparability of other indications

In several instances, the FDA used analogous reasoning about different indications and different drug regimens to support their conclusion. Evidence from randomised trials on other solid organ transplants was used as confirmatory evidence of effectiveness; evidence of safety was equally supplemented from other solid organ transplants; and dosing recommendations were analysed from clinical guidelines and extrapolated from patients receiving heart transplants. Finally, the individual contribution of tacrolimus within the regimen was inferred mostly based on the effect of another CNI medicine regarding kidney transplants.

Generally, the FDA requires two rather than one adequate and well-controlled study to provide 'substantial evidence' for approval. For certain cases, the US Code foresees an exception to this rule if confirmatory evidence is provided. After its initial approval for liver transplantation in 1994, the label of

Prograf had been extended twice already. The first was in 1997, for kidney transplantation, and the second in 2006 for heart transplantation. Although Astellas did not submit any additional evidence or link to previous submissions (Croy, ca. 09:00), the FDA stated that they considered evidence from these RCTs on other solid organ transplant settings as confirmatory evidence of effectiveness. In the multi-discipline review, the FDA referred to the mechanistic comparability of the different indications and reasoned as follows:

> Alloimmune response to these transplanted organs is mechanistically similar, regardless of the organ involved, and rejection is known to occur in the absence of therapy. Therefore, and based on these related uses, it is scientifically reasonable to conclude that tacrolimus as part of an immunosuppressive regimen should decrease and delay rejection in lung transplantation, consistent with the findings of the SRTR study. (CDER 2021, p. 58)

Hence, due to mechanistic comparability between alloimmune responses of different transplanted organs, earlier RCTs were used as confirmatory evidence together with the SRTR study. These studies jointly provided substantial evidence of effectiveness as required by the US law. Similarly, in the scientific publication of this study, the team mentions published clinical trials comparing tacrolimus with cyclosporine; these trials provided additional evidence that was congruent with the team's own results:

> However, the observed rates of rejection at 1 y posttransplant (25.3% in the TAC + MMF group compared with 31.3%–49.4% in the other groups) are in line with the results of multicenter, prospective, randomized trials in lung transplant recipients showing rates of acute rejection at 1 y posttransplant to be numerically lower in TAC + MMF than CsA + MMF groups. (Erdman et al. 2022, p. 1240)

Despite the accepted relevance of the database, it did not contain the necessary data on dosage. Evidence for dosing recommendations was again found in external sources. According to the multi-discipline review, the information came from published 'clinical practice guidelines' other 'published studies' and by 'extrapolation from the dosing information for heart transplantation' (CDER 2021, pp. 14–15). Astellas' lead investigator addressed the FDA's decision as follows:

> Most interesting to me was the agreement to have label recommendations for dose based on published literature, and notably it was primarily from heart transplant patients. ... But that said though, there was a scientific connexion made between a lung and heart transplant to support this, it wasn't just empty. (Croy 14:46)

The review did not state more clearly which scientific connection was made to support the extrapolation of dosing recommendation. The connection that probably served this purpose was a comparison of pharmacokinetics (i.e., the process of absorption, distribution, metabolism and excretion of a drug) between lung and heart transplant patients, which indicated 'similar patterns' (CDER 2021, p. 30).

The historical literature review not only established the mortality rate in the absence of therapy but also aimed to identify the specific contribution of tacrolimus within the regimen-based therapy of lung transplant patients. Tacrolimus was only the second CNI medicine on the market. Hence, the historical information primarily supports the contribution of the earlier product, cyclosporine, to the success of the transplant procedure. Moreover, a crucial piece of empirical evidence referred to a comparison from two trials in the 1990 on kidney rather than lung transplants. In one trial, patients received an immunosuppressive regimen without a CNI; in the other trial, the regimen contained the CNI cyclosporine. Acknowledging the gap between the evidence and the conclusion, the FDA reasons:

> Nonetheless, the similarity of the regimens to those used in the SRTR study as well as the similarity between kidney and lung transplantation are sufficient to support the finding that CNIs contribute to graft and overall survival in the setting of an immunosuppressive regimen in lung transplantation. (CDER 2021, p. 19)

The final piece of evidence was a trial that compared two regimens for lung transplantation, one containing tacrolimus and the other containing cyclosporine; otherwise, the regimens were identical. The evidence suggested similar effectiveness of the two regimens in lung transplant patients. These to studies support jointly the FDA's analogous reasoning that the regimen without tactrolimus would be comparably insufficient in lung transplants.

In summary, mechanistic and analogous reasoning played a crucial role to confirm the effectiveness of the medicine, fill evidence gaps on dosage recommendations and demonstrate the contribution of tacrolimus to the outcome and eventually to support the data as 'fit-for-purpose'. A reference to the mechanistic similarity of alloimmune response in different transplanted organs and the similarity in pharmacokinetic patterns between lung and heart transplant patients is used to justify the analogy. Clearly such contextual evidence has done most of the epistemic work.

5. Fitness-for-purpose: A two-sided sword

On 16 July, the FDA announced that the approval of Prograf for lung transplant recipients reflected that 'a *well-designed*, non-interventional study relying on *fit-for-purpose* real-world data, when compared with a *suitable control*, can be considered *adequate and well-controlled* under FDA regulations' (US Food and Drug Administration 2021a, my emphasis). A closer look at the evidence behind these claims shows that the FDA's confidence in this decision was not solely based on evidence from real-world data. The FDA relied heavily on prior evidence, analogous reasoning and contextual evidence to account for various deficiencies in the data and support its decision.

This case illustrates several of the observations made in the previous sections. The flexibility of the fitness-for-purpose approach enabled the FDA to recognise sufficient evidence for good decision-making where it clearly existed. One might even argue that the FDA's choice of a historical control was a risk-minimising move, since it greatly diminished the risk of falsely denying the approval, compared to a comparative design. Hence, critics who are concerned that the emergence of real-world evidence in the regulatory realm will increase the risks for patients will not find their foil in this case. At the same time, the case raises concerns regarding the future use of real-world data and a contextual approach to data quality. The first concerning aspect about this case is about how many compromises the team made along the way. The research team hedged the primary causal conclusion by accepting treatment effects based on regimens and accepted the uncertainty that accompanies treatment effect estimates that are blind to events occurring during the hospital stay. They also hedged all causal conclusions for secondary outcomes by accepting broad time intervals rather than precise dates. In addition, some research interests could not be pursued because the available data did not allow for them. Most relevantly, to make up for the lack of target trough levels to make dosage recommendations, the team included data from prior evidence and used analogous reasoning to fill this evidence gap. This is clearly a case where the FDA and Astellas used the 'adequacy-given-resources' criteria, where the resources even included prior evidence that could make up for a lack of coverage in the current data.

The second concern about this case is that despite all efforts and many compromises made along the way, the team missed a highly relevant quality issue that rendered the primary analysis uninterpretable. If such data errors occurred in

a clinical trial, there a monitor would probably have caught the issue and corrected it by urging investigators to train their study stuff in clearly distinguishing hospital discharge from hospital transfer when collecting data. It is, however, questionable whether any quality assessment of routine data could have detected the issue. The percentage of erroneous data was also notably low, so it might not have raised the alarm even if the team had tested for its accuracy. The researchers were lucky that the quality issue was clearly visible in the effect estimate of the analysis and the researchers could react accordingly. The example is indicative for the complexity of data quality assessments and risks it can entail. Such limits of retrospective data-quality assessment methods are further discussed in Chapter 5.

Thirdly, the most concerning issue of all is that the FDA simply compensated for this problem by accepting the results of the sensitivity analysis as the primary analysis to be reported in the label. Admittedly, in light of the convincing contextual evidence, the convergence of the sensitivity analysis with the post-hoc analysis, and the plausible reasoning the team applied for their post-hoc analysis, it is indeed reasonable that the FDA made this decision. Nevertheless, such a switch between analyses is a protocol violation that should raise concerns about the possibility of fraud. In addition to the many compromises made before the investigation, the researchers missed a relevant data quality issue and made compromises after the analysis by breaking the rules of the protocol. Chapter 6 elaborates on the problem of trustworthiness of evidence.

The contextualised approach to data quality is a two-sided sword. It can be a powerful tool to increase epistemic opportunities. If used with sufficient skills, expertise and the right normative commitments it is a powerful tool that allows the research community to carefully tailor the choice of methods and data to the community's epistemic and practical aims. The conceptual reflections on the notion show that the notion of fitness-for-purpose involves decisive theoretical and normative commitments including commitments about the necessary epistemic and non-epistemic requirements and the level of informativeness and certainty needed to achieve a purpose. The flexibility of the notion allows to recognise the multidimensionality of 'quality' of evidence, highlight those dimensions that are most essential in a particular context, account for available resources and resource constraints and carefully tailor the choice of methods to these circumstances.

It appears that the contextualised approach is also highly susceptible to misuse and error in absence of the right skills and attitudes. Working with messy data will always yield surprises like the primary result of the Prograf case. Such issues will be the rule and not the exception and dealing with them requires

reflective judgments and the right attitudes. Prograf, however, is the exception and not the rule. The amount of contextual evidence available in this case is extremely rare. Most diseases do not have such high mortality, most treatments are not as effective. Moreover, the intense use of such analogous reasoning is puzzling. The reason that indication extensions generally require the same evidence requirements as new approvals lies precisely in non-comparability. As my analysis has shown contextual evidence has done most of the epistemic work in the case of Prograf, even though the SRTR database is arguably one of the better real-world data sources. This point raises concerns about how much epistemic work even high-quality real-world data could do by itself in the absence of such contextual evidence. Moreover, it appears that the team was merely lucky that the quality issue in their data was so clearly visible in the effect estimates as imprecision and not a bias so they could react accordingly. In future cases such a quality issue might well go unnoticed. This raises more concerns about the risks future uses of real-world data will bring.

Chapter 5
The power and reliability of real-world data

The internationally renowned medical newspaper STAT called it 'one of the most seductive ideas in medicine that 'real-world evidence,' including data from electronic health record systems and even records of insurance payouts, could replace the far more expensive and time-consuming studies currently considered the gold standard' (Herper 2019). Two years earlier, the 21st Century Cures Act had introduced the possibility of using real-world evidence for the approval of medicines as an alternative to RCTs and had mandated the FDA to evaluate this idea. The idea has put a global movement into motion. Following the FDA's initiative, the EMA announced in November 2021 that they envisioned enabling the use of real-world evidence and establishing its value for regulatory decision-making in Europe by 2025 (Arlett et al. 2021). The Swiss regulatory agency, Swissmedic, published a brief position paper in July 2022 about their openness to consider real-world evidence in certain circumstances (Swissmedic 2022). The globally recognised reference organisation for evidence standards in clinical research, the ICH, is currently revising two of its core guidelines to include innovations from the real-world evidence movement.

From the beginning, opinions on the amendment diverged. Advocates argued that it would accelerate and multiply therapeutic opportunities for patients. Opponents warned that the easing of evidential standards would expose patients to unnecessary risks. The main motivation behind the evolution of evidence standards is the need for pharmaceutical innovation and concerns about the high costs of drug development. Wilholt has drawn philosophers' attention to the problem that resources matter. In a resource and time constraint research environment, evidence standards must balance the need for certainty with the risk of ignorance. It is thus epistemically desirable to rely on methods that produce results more efficiently. Wilholt calls this desideratum a method's 'power' (Wilholt 2013, 2016). Real-world data is not only a powerful idea; it is also the epitome of a powerful method in Wilholt's terms. The promise of real-world data is not just that it is quickly available as a cheap by-product of healthcare but also that we can reuse

the data to answer many different questions. Wilholt's insight, however, holds that there is always a trade-off between power and reliability. In light of this notion, the ongoing evolution of evidence standards can be seen as a shift in the established power–reliability trade-off, in favour of power and at the cost of reliability.

This chapter is dedicated to an epistemic study of both, the power and the reliability (or quality) of real-world data. My goal is to challenge whether real-world data can fulfil explicitly or implicitly articulated expectations about its reliability and its power. Regarding data reliability, I focus my critique on the problem that the methods used to assess the reliability of data tend themselves to be unreliable. This point is particularly relevant if a threshold already makes a critical allowance for decreased reliability, as in the 'fitness-for-purpose' approach to data quality. If thresholds make an explicit allowance for reduced reliability, the techniques we use to establish the reliability of data should clearly not introduce additional uncertainty. The strength of my argument is that it applies to any normative reliability threshold. For this discussion, I engage in an in-depth epistemic study of data-quality assessments to establish the fitness-for-purpose of data. I criticise two approaches. The first is the idea that purpose-independent data-quality assessments can be performed and standardised on the side of the data provider and used as a reliable indicator of purpose-specific data quality. Against this trend, I argue that not only the thresholds of data-quality measures but the metrics themselves need to be purpose-specific to be reliable. The second trend is the validation of data against an external reference standard. Combining three case examples, I show that three commonly used reference standards all lack epistemic rigour and are therefore generally unreliable standards to quantify data quality.

The second and shorter part of the chapter focuses on the data's power. I argue that real-world data is not well suited to deal with time or resource constraints. Regarding the problem of time constraints, I argue that real-world data can only avoid these constraints in retrospective contexts where we typically lack a true unmet medical need. For the problem of resource constraints, I argue that real-world data might not reduce but mostly distribute the costs for the production of evidence towards public third parties. These could include governments, healthcare facilities, regulators and patients.

Here is how I proceed: Section 1 substantiates the need for fast and cheap data using Wilholt's notion of power. I introduce two different regulatory approaches to articulate implicit and explicit assumptions about the reliability and power of real-world data. Section 2 examines the reliability of data-quality assessments. To illustrate in detail how such an assessment could look like, I first

provide a complementary analysis of the case study from Chapter 4 and explore the FDA's reasons to accept the SRTR as 'fit-for-purpose' to support the effectiveness of Prograf (section 2.1). I then examine the methods used for quality assessments from two angles. The first concerns purpose-independent quality assessment on the side of the data provider (section 2.2). The second angle is the validation of data against an external reference standard (section 2.3). Section 3 turns to the perceived power of real-world data. I first criticise the assumption that real-world data can deal better than randomised trials with time constraints (section 3.1.). In the final section, I examine the assumption that real-world data can deal better with resource constraints (section 3.2).

1. Costs of clinical trials and power of real-world data

Developing new drugs is an expensive and risky business. Recent empirical research has estimated that the median costs for bringing a new compound to the market is around $320 million. If additionally the costs of other compounds that failed during the development phases are factored in, the median costs exceed $1,1 billion (Wouters et al. 2020). Earlier estimates reached comparable figures by estimating the development costs at $800 million, of which 30% to 60% are spent entirely on the clinical development (Rawlins 2004, p. 360). Others have estimated that for medicines that reach market approval, about 90% of all expenditure relates to phase III clinical trials (Roy 2012). Furthermore, while the costs of bringing a compound to the market have grown vastly in recent decades, the approval rate for new medicines has remained stable since the 1950s (Munos 2009).

A commonly cited reason for the steady increase of costs is the steadily increasing regulatory requirements. Munos estimates that regulatory requirements cause a yearly increase in the costs of more than 8% (Munos 2009). Others have found that about 50% of the clinical trial budget is allocated to activities to ensure compliance with the GCP guidelines, of which 50% is allocated to the on-site verification of data by monitors alone (Funning et al. 2009). There is considerable pressure on the industry, regulators and governments to decrease the costs of the overall development process and the regulatory requirements in particular. It has even been argued that the increased complexity of the regulatory requirements has made pharmaceutical companies dependent on so-called contract research organisations, which provide professional but expensive research services to these companies (Collins et al. 2020). Rawlins has called for the rigorous examination of

the 'rituals' in the drug development practices and to test these regulations for their evidence basis and their cost-effectiveness (Rawlins 2004, p. 361).

These concerns are not only coming from the industry but also from regulators themselves. Historians and social scientists have argued that regulators see themselves not only as gatekeepers but also as enablers of innovation (Hauray 2017). A pressing issue for regulators was described by former senior medical officer at the EMA, Hans-Georg Eichler, as the 'opportunity costs' of high regulatory standards (Eichler et al. 2013). In his view, regulators are generally quite risk-averse and have a low tolerance for uncertainty. Decreasing uncertainty and meeting high standards in one case requires valuable resources that will no longer be available for another research undertaking. In Eichler's words:

> The issue at stake here is this: if the resources required for the second trial had been reallocated to another drug development programme, how much more overall knowledge — and ultimately health benefit — could have been gained? (Eichler et al. 2013, p. 911)

Particularly in times where basic research produces more potential hypotheses than can be pursued with available resources, Eichler argues, all resources spent on one claim create an opportunity cost for all other unpursued hypotheses. Eichler follows Roy in embracing the bold view that opportunity costs are even plausibly responsible for the 'great tragedy of the pharmaceutical industry'. He describes this tragedy as the problem of 'promising drugs that are not being prescribed because of the expense and risk of developing them' (Roy 2012, p. 7). Eichler speculates that opportunity costs could explain the disinterest of the industry in antibiotics (Eichler et al. 2013), while Roy cites the industry's disinterest in common illnesses such as heart disease, stroke and obesity (Roy 2012). Eichler and colleagues have argued that regulators should be allowed to factor such opportunity costs into the development of evidence standards and decisions about individual cases.

Academic clinical researchers have also emphasised that routine data and pragmatic trials are valuable because they provide cost-effective evidence. Hemkens and colleagues note that it is unlikely to approach an 'exhaustive evaluation' of treatment effects and harms with RCTs and advocate that less costly real-world studies could fill many evidence gaps (Hemkens et al. 2016, E159). Mc Cord and colleagues similarly argue that real-world evidence can help extend the research agenda to questions that are not generally approached by the industry (Mc Cord et al. 2018).

In a resource-constrained research environment, general concerns about the correct allocation and efficient use of resources are well justified. Wilholt's notion

of a method's 'power' substantiates these concerns philosophically (Wilholt 2013, 2016). Traditionally, evidence standards have been seen as a trade-off between two types of reliability, namely a type I error (the error of accepting a claim that is in fact false) and a type II error (the error of rejecting a claim that is in fact true). The insight that the right balance cannot be determined in a purely epistemic way but depends on the social and ethical consequences of making either of these errors has prompted fruitful research on the roles of non-epistemic values in research. The most famous example is the work of Heather Douglas (Douglas 2000; Douglas 2009). Wilholt has convincingly argued that evidence standards also balance an overlooked third risk, namely the risk of not obtaining any result within the available time and resources (Wilholt 2013). Consequently, methods are not only epistemically valuable in terms of how likely they are to produce a true result but also in terms of how efficient they are in producing any result at all. In a world with potentially endless scientific questions but limited time and resources, this epistemic desideratum of a method matters to our epistemic aim of accumulating scientific knowledge. We not only want reliable knowledge; we also want as much knowledge as possible. Wilholt referred to the 'power of the method' to describe this epistemic desideratum. Following Wilholt, we define the power of a method as: 'the rate at which a method or type of inquiry generates definitive results, given a certain amount of effort and resources' (Wilholt 2016, p. 227). In Wilholt's definition, this rate includes all conclusive results produced by a method, whether they are true or false. As he notes, the difference between negative results and inconclusive results is not always clear-cut. Particularly, experiments that lead to results below the statistical significance threshold can count as either negative or inconclusive, depending on the statistical theory used or the applied statistical practice. Cases where experiments are discontinued or not even started because of resource or time constraints count as failing to produce any result. Crucially, in clinical research, resources are not only a matter of money but also of patients who are available to become research participants. Empirical studies have shown that the failure to recruit sufficient patients is the most common reason for discontinuation of a clinical trial in academic research (Briel et al. 2016).

Following Wilholt, the epistemic power of a method generally trades-off with its reliability, where reliability means the risk of making a traditional type I or type II error. Reliability could always be increased, he argues, by increasing the amount of evidence that is required to support a particular claim. For example, instead of the common rule of requiring two clinical trials for market approval, regulatory agencies could demand five such trials. The generally accepted

significance threshold of 0.05 could be increased to a threshold of 0.01. Clearly, more resources are required to meet these higher standards, and there is nothing to gain from such an increase if we lack the resources to uphold these standards. Hence, concerns for power are what put an end to the potentially unlimited increase of the reliability of our claims. Consequently, in Wilholt's view, settling on conventions for evidence standards is not a convergence on the most reliable standards but rather a conventional agreement about how to balance power and reliability. To illustrate, Wilholt contrasts the preference for randomised trials over observational studies in EBM. In his view, the question which method is more suited to scientific inquiry 'makes no sense' because they balance power and reliability in different ways (Wilholt 2016, p. 227). Whereas randomised trials are better at answering a few questions more reliably, observational studies are better at answering more questions with the same resources. Finding the right balance for evidential standards is therefore a question of making the right value judgements. Evolving evidence standards are a consequence of evolving preferences about ethical or social goods and 'irreducibly social' (Wilholt 2016, p. 219). The current evolution of evidence standards can be seen as a shift in the established power–reliability trade-off, in favour of power and at the cost of reliability.

Critics of the current regulatory standards, such as Eichler and colleagues, need to respond to the criticism that their proposed evidence standards do not get this balance right or that they focus on the wrong ethical and social goods. For example, in Fraile, Tempini and Teira's view, the new standards might wrongly prioritise the freedom of patients over impartiality (Fraile Navarro et al. 2021). Stegenga has argued that we should care more about cost-effectiveness of treatments rather than costs per se (Stegenga 2017). My goal here is not a critique of the normative balance. My goal is to challenge whether real-world data can fulfil explicitly or implicitly articulated expectations about the epistemic and normative goods that can be gained, and the losses in reliability that are paid, regardless of how well these normative goods are justified ethically or socially. Hence, I am interested in these two questions: Can real-world data deliver the expected gain in power and the expected normative goods? Can the data meet expectations about its reliability? To apply to data, I interpret reliability in terms of quality dimensions like accuracy completeness of data. If data tend to be less reliable than expected, the evidence generated with them risks being less reliable as well.

To begin, I roughly establish what the expectations of stakeholders in the field are regarding the data's power and reliability. There are two kinds of regulatory approaches that allow for using real-world evidence, which articulate

different expectations and balance the power and reliability in different ways. The first regulatory approach is the use of real-world data in accelerated approval programmes or related programmes that explicitly foresee an exception from the standard requirements for evidence. In the past, the FDA has adopted a wide range of programmes that knowingly and explicitly accept the premise that the evidence does not fulfil their criteria for 'substantial evidence' at the time where patients get access to the treatment. In return, patients gain earlier access to potentially life-saving treatments. In this trade-off, the increased power accelerates the approval process.

A well-known example of this regulatory approach is the adoption of so-called compassionate use programmes, where patients can get access to treatments that are still under development. Another example is the EMAs 'adaptive pathways' programme, launched in 2014, in which selected treatments receive a conditional or restricted market authorisation based on preliminary evidence, and real-world evidence is used to quickly reduce the uncertainty and adapt decisions accordingly (Davis et al. 2016; European Medicines Agency 2014). Gloy et al. demonstrate that the FDA has made wide use of this option for new cancer treatments during the last 20 years. This has led to the situation of only 7% of newly approved cancer drugs having been supported by the conventional two clinical trials; all the others were supported by only one randomised trial or none at all (Gloy et al. 2023). Fraile Navarro and colleagues describe this driver for the change in evidential standards as a trade-off between impartiality and the freedom of patients (Fraile Navarro et al. 2021). The applicability of these programmes is restricted to treatments and diseases that fulfil certain ethical or social requirements to count as high-priority among regulatory agencies. Stegenga has cast doubt on the assumption that these programmes deliver on their expectations by arguing that the new medicines are barely the life-saving good that people hope for (Stegenga 2017). Others have argued that the EMA's adaptive pathway programme cannot fulfil the expectations about its promised reliability (Davis et al. 2016). In section 3 of this chapter, I support the critics by arguing that the most essential time gain with real-world data can only be expected for treatments whose ethical benefits are marginal.

The FDA's real-world evidence programme is part of a second regulatory approach for using real-world data. Here, the use of the data is not restricted to particularly valuable treatments and it does not explicitly sell the current reliability standards for an ethical or social good. Instead it aims at using real-world evidence more generally, while maintaining similar evidence standards for decision-making. For example, the US Code on the Utilization of Real-World Evidence states that

allowing the use of such evidence under this article should not be interpreted as undermining the 'substantial evidence' and 'adequate and well-controlled' evidence requirements (US Code, Title 21, Chapter 9, Subchapter 5, §355g). However, this regulatory proposal still allows for a reduction in the reliability of data. This reduction in reliability is made explicit in the notion of 'fitness-for-purpose'. The aim of this approach is to identify data that is *sufficiently* accurate to support good decision-making. (For a discussion of the fitness-for-purpose approach to data quality, see Chapter 4).

The underlying assumption of this regulatory framework seems to be that the current evidence standards are, at least in some instances, either more reliable than necessary – or ineffective at promoting reliability. What this means is clearly illustrated by the second revision of the GCP Guidelines in 2016. A cornerstone of these guidelines is the mandatory verification of clinical research data by so-called monitors. This 'source data verification' practice was established in 1988 by the FDA's guideline on Monitoring of Clinical Investigations. The goal was to achieve 100% accuracy and completeness of data submitted to the FDA, relative to the respective source data (Andersen et al. 2015). Empirical evidence indicates that this practice is effective and reduces the error rate to as little as 0.0% – 0.36%, depending on the type of data. Nonetheless, in 2011 the FDA withdrew its guidance in favour of a so-called risk-based approach to monitoring. In 2013, the EMA followed suit with their reflection paper on risk-based monitoring for clinical trials. In 2016, the second revision of the GCP added an Addendum, which similarly recommended a risk-based approach to monitoring (International Council for Harmonisation 2016; Andersen et al. 2015). The idea behind this shift is that we do not need error rates that are as low as 0.36% for all types of data in a clinical trial in order to license valid, precise and statistically significant causal inferences. Monitors should therefore focus on trials and data items that pose a high risk to the integrity of data or the safety of participants and should thus assess the data 'proportional' to the risks they pose. In the second revision of the GCP guidelines, the purpose of these revisions and the problem are clearly addressed:

> Since the development of the ICH GCP Guideline, the scale, complexity, and cost of clinical trials have increased. Evolutions in technology and risk management processes offer new opportunities to *increase efficiency* and focus on relevant activities. ... Therefore, this guideline has been amended to encourage implementation of improved and *more efficient approaches* to clinical trial design, conduct, oversight, recording and reporting while continuing to ensure human subject protection and reliability of trial results. (International Council for Harmonisation 2016, my emphasis)

The idea of proportionality introduced in the second revision of the GCP guidelines was reinforced and expanded in the third revision of the guidelines. The third edition introduces the idea that the overall quality of a clinical study is its 'fitness-for-purpose'. Furthermore, high-quality designs attend proportionally to those aspects of the study that are 'critical to quality factors' (International Council for Harmonisation 2023, 2021b). In this proposal, the increase in power aims at decreasing the material resources required to achieve a result. Its use is not restricted to medicines promising a high social good. However, the reliability of the data should be sufficient for good decision making, that is, data should be 'fit-for-purpose'. The two proposals highlight three expectations about the data's power and reliability: accelerated access to priority medicines, decreased costs of drug development and sufficient reliability for good decision making. The following sections analyse whether real-world data can deliver on these expectations, beginning with a critical study of the reliability of real-world data.

2. Critical study of the reliability of real-world data

Real-world data is acceptable as evidence if it is 'fit-for-purpose'. A conceptual discussion of this notion in Chapter 4 has shown that putting the notion to use requires defining the purpose and the criteria required to achieve the purpose. These generally include properties such as accuracy, completeness, or relevance of data but determining the necessary properties depends on the purpose of the study. Here, I subsume the list of these properties under the idea of reliability. That is, data are reliable if they are accurate, complete, relevant, or other required properties. To clarify, the same argument could be cast in terms of quality rather than reliability and sometimes I use the terms interchangeably in this chapter.

The notion of fitness-for-purpose makes an allowance for data to be less than perfectly reliable, as long as the data is *sufficiently* reliable for sound decision-making. Directly targeting the idea of sufficient reliability is challenging. Since the standard makes an allowance for reduced reliability, simply pointing out discrepancies from the current standard is an idle strategy. Other philosophers have answered the challenge by gesturing towards the concern that real-world data might not be just a little less reliable but rather a lot less (John 2021). We might read such a gesture as doubting that real-world data can in fact be fit-for-purpose given a certain normative idea about what good decision-making entails. I support the critics, but my strategy is a different one. I accept the idea that real-world data

can be fit-for-purpose to support decisions about the effectiveness of medicines. Yet, I doubt that we can reliably assess whether some data is in fact fit-for-purpose in a particular instance. Hence, I do not target the normative reliability threshold but rather the methods and strategies that are used to establish whether data meets this threshold. Here I am interested in the methods and practices that are used to determine that data is *sufficiently* reliable to be used for a particular purpose, that is, to determine whether data is 'fit-for-purpose'. If this argument is successful, it follows that real-world data is not acceptable as evidence because we cannot reliably establish that it is fit-for-purpose.

The difficulty with this strategy is that the methods and techniques used to assess the quality of data are manifold, and discussion of such methods is still relatively scarce in the philosophy of science. This undertaking therefore requires a detailed epistemic study of these techniques – which accounts for the length of this chapter. Section 2.1 provides a starting point and a guiding example by taking a closer look at the FDA's quality assessment of the SRTR, which led to the FDA's first approval exclusively based on real-world evidence in 2021. I then turn my attention to attempts to systematise data quality frameworks and standardise data quality checks on the side of data providers (section 2.2). This is followed by a longer discussion of the practice of validating data against an external reference standard (section 2.3). Both techniques were used in the assessment of the SRTR and are repeatedly recommended or envisaged as reliable techniques in regulatory guidance on data quality.

2.1. Assessing the quality of the SRTR

In July 2021, the FDA approved the use of Prograf as an immunosuppressant for lung transplant recipients exclusively based on real-world data from the SRTR. The recent label extension was the third extension after its initial approval for liver transplant recipients and later for kidney and heart transplants. Chapter 4 contains an in-depth discussion of this decision and the overall evidence submitted to the FDA. I argued there that the data submitted to the FDA can be regarded as adequate and well-controlled in light of the highly conclusive contextual evidence, such as the large effect size and comparability of indications. Here I complement the analysis of this case and I explore why the quality assessment the FDA and Astellas used in this case were insufficient to catch a particular problem in the data. I also examine what this failure can teach us about data quality assessments in general.

To recall, the database used for the study was the US SRTR. Since 1987, the SRTR collected data on all solid organ transplants in the US, by law. The database has been used primarily for the purpose of facilitating the regulation of organ donation and allocation in the US. It allows registering of candidates for transplants, matching of donated organs to candidates and submission of data on donors, candidates, and recipients for and after transplants. Moreover, the law required the mandatory and comprehensive reporting of transplant procedures and outcomes into the registry (For details see Chapter 4).

I now take a closer look at the methods used to determine that the SRTR data is fit-for-use. According to the multi-discipline review by the FDA, their assessment of the data as being fit-for-purpose was based on a wide variety of criteria and assessment methods (CDER 2021). In summary, these are the reasons that appear in the FDA review:

1) High regulatory oversight and mandatory data collection for all US solid organ transplants.
2) The database has a well-established and robust operational structure.
3) The processing steps within the database include the verification of data quality.
4) Some data, including mortality data, is verified against data from *other trusted sources.*
5) An assessment of missing data revealed no concerns.

To support their judgement for 1)–3), the FDA relied on a ten-year-old publication by Leppke et al. describing the legal basis, operational structure, data collection and data use of the SRTR in some depth (Leppke et al. 2013). According to Leppke and colleagues, the SRTR is indeed embedded in well-established legal and operational structure. The SRTR is a database with the main purpose of facilitating organ procurement and allocation in the US and improving policies on which these decisions are based. Its legal basis is the National Organ Transplantation Act, passed into law in 1984. By its mandate, the registry collects data 'necessary to an ongoing evaluation of the scientific and clinical status of organ transplantation', which has been mandatory for all US solid-organ transplant recipients since 1987 (Leppke et al. 2013). The National Organ Transplantation Act simultaneously mandated the establishment of the OPTN to overview a national strategy for organ matching. The contracts for both mandates (OPTN and SRTR) are managed by the Health Resources and Services Administration within the US Department of Health and Human Services and overseen by the division of transplantation. Since its foundation, the SRTR has been operated by three different institutions. Since 2010 the operation of the database has been mandated to the Hennepin Healthcare Research Institute (formerly Minneapolis Medical

Research Foundation) and executed by their CDRG. The SRTR is further overseen by a steering committee and a technical advisory committee. The director oversees the SRTR senior staff, including 18 experts from diverse fields, such as clinical experts and biostatisticians (Leppke et al. 2013).

Regarding data collection and processing steps, most of the data in the SRTR is collected by the OPTN. They collect data from transplant centres, organ procurement organisations and histocompatibility laboratories (Leppke et al. 2013). The data collection is guided by standardised data collection forms and OPTN policies on data collection. These policies define which entity must submit what types of data within what specific timeframe (Organ Procurement and Transplantation Network 2023). To further check the quality of this data, the SRTR applies several data processing steps, including the validation of data quality. Importantly, these measures include a feedback loop with the data providers to address and resolve data quality issues. Leppke et al. stated that

> Numerous tests are performed to identify potential errors in the data. A report is generated and sent to OPTN so these issues can be addressed and resolved, enhancing the overall quality of the data in future releases. (Leppke et al. 2013, p. 53)

The authors do not describe in detail what type of tests are applied or what exactly these tests can support.

The operational structure of the database is of relevance for the FDA's assessment because the processes through which data are generated and made available causally determine several dimensions of data quality like their accuracy or their timeliness. Leonelli aptly writes that 'data are regarded as reliable on the basis of the methods, instruments, commitments, values and goals employed by the people who generate them' (Leonelli 2017b, p. 3). The EMA's data quality framework calls these the 'foundational determinants of data quality' and clarifies that assessing them is intended to ensure that the data is collected and curated in a way that the 'correspondence between the data and the real entity is not altered' (European Medicines Agency 2022, p. 9). In this example, the fact that the SRTR at least has policies for the timely collection of data speaks in favour of data quality, because it is likely that data is in fact collected according to these policies. I discuss the relevance of these criteria in the next section 2.2 on purpose-independent quality assessment.

What seems to be of greater relevance in the FDA's review than the SRTR's foundational drivers of data quality is reason 4), namely the database's linkage to other data sources that the FDA deems 'trusted'. The first external data source is

the Social Security Administration's Death Master File. This is the most comprehensive list of all diseased individuals in the US. According to Leppke et al., this data is important because the OPTN policies do not require any follow-up on transplant donors and candidates, only on recipients, beyond two years. Additionally, it seems that the SRTR also uses data from the Death Master File for complete verification of the data on death among organ recipients. That is, the data from the Death Master File is used as a reference standard to complete and correct the data for each individual patient.

The SRTR is also enriched by data from the Centres for Medicare and Medicaid regarding patients with end-stage renal disease. Not only the FDA, but also the research team, affirmed their confidence in the assessment of the primary outcome measuring death or graft failure because the SRTR relies on these external data sources. The high level of confidence in this data item in the SRTR is expressed by other experts too (Stock 2017, p. 3001). Within the FDA's multi-discipline review, there was no overt justification for their reliance on these other data sources; however, the Social Security Death Master File is generally known and acknowledged as the most reliable source on deaths in the US. This approach to data quality is covered in the discussion of data validation in section 2.3.

Finally, the team took additional steps to assess and increase the quality of the data at the level of the individual study, as stated in item 5). They assessed the extent and impact of missing data by measuring the percentage of patients who had complete data for all secondary outcomes. They found that the missing data amounted to 5% or less, except for one outcome. To assess the impact of the missing data, they performed a sensitivity analysis, where they assumed that all cases with missing data would in fact be cases that experienced the primary outcome (death of graft failure). They concluded that 'data availability was considered relatively robust'. They acknowledged that missing data in the treatment arm of their main interest (rather than in the overall population) was higher than 5% and could be as high as 14.2%; however, they still concluded that 'Data availability was considered relatively robust in this study' (Erdman et al. 2022, p. 1241). Another measure they took was to exclude patients with data patterns that were unsuitable to answer the study question. This step involved the exclusion of about 6% of patients who died during their stay at the hospital, to prevent immortal time bias. A further 3.3% of patients were excluded due to missing data on their immunosuppressive regimen at discharge (Erdman et al. 2022).

Overall, the FDA's assessment of the data as being fit-for-purpose is based on a wide variety of indicators and assessment methods in line with the respective

guidelines on real-world data (US Food and Drug Administration 2021c). Yet, all these measures did not prevent the team from running into a severe problem with their primary analysis because of a data quality problem. To recall, here is brief version of what went wrong: The team did not discover the problem that a small percentage of the data reported implausible early hospital discharge dates. Patients who experienced an event (death or graft failure) and were discharged from the hospital were considered by the analysis and unfortunately the team chose an analysis method that was highly sensitive to early events. Most likely, the early hospital discharge dates are an error in the data and occurred because some sites did not properly distinguish between hospital discharge and hospital transfer (PSI RWD SIG 2021; CDER 2021).

The example is indicative for the complexity of data-quality assessments and the risks it could entail in other circumstances. Luckily, the quality issue was clearly visible in the effect estimate of the analysis and the researchers could react accordingly. In the following sections I examine two practices for data quality assessments more systematically to see what the case of Prograf implies for the reliability of data quality assessments in general.

2.2. Purpose-independent quality checks by data providers

A critical challenge for data quality practices is to bridge the gap between the highly abstract success criteria, 'fitness-for-purpose' and concrete quantitative assessments of data quality. To that end, quality frameworks usually disentangle the notion into various quality dimensions. The difficulties of this endeavour are discussed in Chapter 4. The second important aspect of data quality assessments is the definition of metrics and concrete quality checks that are performed to support claims about data quality.

To illustrate what such metrics could look like, here are a few examples from the EMA's data quality framework to assess the quality dimension called 'plausibility':

a) Height and weight are positive values
b) Discharge date happens after admission date
c) Sex values agree with sex-specific context, such as for prostate cancer
d) Oral and auxiliary temperature for the same patient agree
e) Recorded date of birth is consistent between EHR data and registry data for the same patient
f) Count of immunisation per month shows an expected spike during flu season. (European Medicines Agency 2022)

It goes without saying that the possibilities to construct such data quality metrics are endless. However, following Illari, these metrics have an interesting property: Unlike the abstract notion of fitness-for-purpose, which is deeply purpose-dependent and can only be understood in the context of a specific use, such metrics can apparently be constructed and applied to data regardless of the data's context of use. Hence, unlike the notion of fitness-for-purpose, quality metrics are not necessarily a relational property of the data and its user. The metrics can be sensibly defined according to the data itself; i.e., they are properties of the data (Illari 2014). For Illari, this property of quality checks provides a path out of the problem which she calls the 'rock and the hard place' problem. Data-quality practices are between 'a rock and a hard place' because data sharing requires making (high-quality) data available in a purpose-agnostic manner, whereas data quality practices are limited if the data's purpose is unknown. (Illari 2014). In other words, the purpose-independency of quality metrics opens the possibility to perform purpose-agnostic quality checks on data sets by the data providers and data curators. The provider can then report about the success of these checks as a quality indicator. Reason number 3) cited for the assessment of the SRTR database precisely relied on such purpose-agnostic 'verification' of the data by the data providers.

It seems already to be common practice that data infrastructures perform multiple such quality checks to support the quality of their data. Callahan and colleagues examined data quality practices across six large data-sharing networks in Europe; they compared the infrastructures by counting the number of quality checks performed and mapping them onto the quality dimensions by Kahn et. al. They found that these infrastructures performed between 174 and 3434 quality checks on their data, most of which classify as plausibility checks (60%), followed by conformance checks (27%) and completeness checks (13%) (Callahan et al. 2017). In the case of the SRTR, a naïve version of this approach was used when the team relied on the data providers' general report that data is verified by internal processes. The publication by Leppke et al. on which the FDA relied did not further specify what precisely has been verified (Leppke et al. 2013). The less naïve version is what the EMA envisages for the future, when data infrastructures are likely to develop into more mature organisations. A standardised set of quality checks will be performed across datasets and infrastructures and packaged as metadata to accompany the data on its journey to the user (European Medicines Agency 2022).

The problem I see with this approach is that data quality metrics are not truly purpose-independent. Such quality checks by data providers are therefore

insufficient to establish the purpose-specific quality of data. To begin with, the different metrics clearly differ in how permissive or strict they are for what counts as plausible data. Consider metric a) 'Height and weight are positive values' and b) 'Discharge date happens after admission date'. Both are weak criteria for plausibility, because the requirement is also satisfied by a wide range of otherwise implausible values. Metric f), which compares summary statistics with medical background knowledge, is an even weaker indicator of the plausibility of individual-level data; again, a wide range of implausible or inaccurate values would fulfil this requirement. Finally, e), which assesses the consistency between different data sources, depends on the accuracy or plausibility of the external source. If such a source can be considered a gold standard, consistency with this source strongly confirms the accuracy of individual data values. But if it is simply another independent data source, agreement between the data values would provide some – but certainly weaker – support of plausibility.

Although the number of quality checks performed by data infrastructures is impressive, the endless possibilities of such checks together with their unequal support for data quality implies that no amount of purpose-agnostic quality checks can provide sufficient support for the plausibility of any particular subset of the data for a specific context of use; only the right quality checks can. For example, if a metric of the form b) 'Discharge date happens after admission date' was used to assess the SRTR data, 100% of the SRTR data would have passed this plausibility check. By contrast, a plausibility check of the form b) 'Discharge date happens after the minimally plausible length of hospital stay' could have revealed the problem in the data.

Of course, one could package more information into the metadata and report the precise metrics employed to assess the data quality. This unburdens researchers to perform the tests themselves, however this does not change the fact that a contextualised, and skilled interpretation of these measures is required to see whether they are sufficient for a particular purpose. Following a lesson taught by Leonelli, interpreting metadata remains a complicated task which requires experts skills (Leonelli 2017b). Interpreting data quality checks is not just a matter of setting the right quantitative threshold required for data to be fit-for-use but rather a matter of defining the right metrics that are informative for the kind of plausibility required for a specific purpose. Hence, the purpose-dependent nature of data quality persists even at the level of concrete data quality metrics, despite their seeming to be purpose-independent. Consequently, no amount of quality metrics is a reliable

indicator of the quality of any particular subset of data, unless one ensures that the correct metrics are included.

Reason 2) on the FDA's assessment of the SRTR, namely 'the database has a well-established and robust operational structure' suffers from the same problem. Assessing the operational structure of a data infrastructure is relevant for the quality of data. It might convince data users to use some data as a *starting point* for their research because such data could be of sufficient quality (Leonelli 2017b). However, without assessing whether the operational structure is sufficiently robust for the precise purpose of the study, the purpose-independent assessment thereof is not a reliable indicator for the kind of data quality required by the study. To prevent the problem in the SRTR data, we require specific policies and operational structures that are quality-conducive for the type of data required in the study.

The issues I raise here are not a principle concern that no data quality checks could have caught the problem in the SRTR data. In principle, the team could easily have subjected the data to the purpose-specific quality check that asks whether 'Discharge date happens after the minimally plausible length of hospital stay' and they might have recognised the problem in the data. My argument points out that there is no escaping the contextuality of such an approach, which requires skilful and purpose-specific engagement with quality metrics. Hence the failure of the team to capture the problem in the data nevertheless points to another general concern regarding the broad fitness-for-purpose approach to data quality. A plausible reason why they did not perform such a test is that they might not have known about the impact that discharge dates could have on the analysis; after all, it seems not a straightforward relation. Following their own reports, it was only after they were surprised by the results that they investigated the reasons for the problem and found that the analysis they had chosen was highly sensitive to these early events (PSI RWD SIG 2021). This implies that they simply failed to identify erroneous discharge data as critical to the quality of the data before the study and therefore did not think about testing the data's plausibility. It is precisely because data quality assessment is such a highly contextualised undertaking that it requires, at any and all points, knowledgeable, skilful, laborious and sincere handling by experts. This is, of course, the lesson that Leonelli has taught philosophers some time ago (Leonelli 2016). The contextualised approach to data quality assessments raises the risk that critical details might go unnoticed, as they did in the SRTR case. The main problem here is not one of reliability of the methods but rather of the trustworthiness of data quality assessment, which I discuss in the final chapter

(Chapter 6). For now, I conclude that a purpose-independent assessment of data quality is an unreliable indicator of the fitness-of-purpose of a particular dataset.

2.3. Data validation against an external standard

The FDA guideline on using EHR data for submission to the FDA seems to suggest that the preferred and most reliable method for data quality assessment is – what the FDA calls – 'complete verification' of data. By this the FDA refers to a procedure where all individual subject–level values are compared against an external trusted gold standard and corrected accordingly. This is precisely what the SRTR does with data on mortality, by comparing it against data from the Death Master File – which the FDA deems a 'trusted data source'. Established gold standards that allow for the complete verification of data are hard to find. A related procedure is the validation of a subset of data against a reference standard, which is then extrapolated to describe the accuracy of the dataset in terms of measures known from diagnostic tests, such as sensitivity or positive predictive value[20]. Following the FDA's non-binding recommendations, this procedure is considered the second most reliable technique and has been repeatedly recommended (US Food and Drug Administration 2021b). The same recommendation is made by the NIH Pragmatic Trials Collaboratory (NIH Pragmatic Trials Collaboratory 2014).

In this section, I engage in a critical study of the method and explore how reliable such validation procedures are. My main critique is easily summarised: the validation of data against an external reference standard is as reliable as the reference standard used. As I show with three examples, the most commonly used reference standards for such studies encounter various problems.

The general approach of data validation against an external reference standard includes three broad steps (A summary is provided in Figure 2):

1. Classification of the data that needs to be validated
2. Construction of a reference standard dataset
3. Quantification of the data quality against the reference standard

I elaborate on the first step using my first example. Researchers aimed to evaluate different algorithms for classifying acute exacerbation in COPD patients in the

[20] Different data quality frameworks use the terms verification and validation differently. Most popular became the distinction introduced by Kahn et al. In their framework 'verification' refers to the comparison of data against 'internal' constraints (such as background knowledge, consistency relations). They use the term 'validation' to refer to practices that compare data against an external reference standard. Kahn et al. (2016).

Clinical Practice Research Datalink (CPRD). The study was performed by Rothnie and colleagues (Rothnie et al. 2016). In 2015 the database contained UK general practice data from about 4.4 million registered patients, or 7% of the UK population; it has reportedly been used in about 1000 studies (Herrett et al. 2015). Acute exacerbation of COPD is an outcome that is often used in real-world studies, as it is robust in the sense that it usually requires patients to contact the care facility to receive appropriate treatment. The comparative frequency of acute exacerbations can then be used to estimate the effectiveness of new COPD treatments.

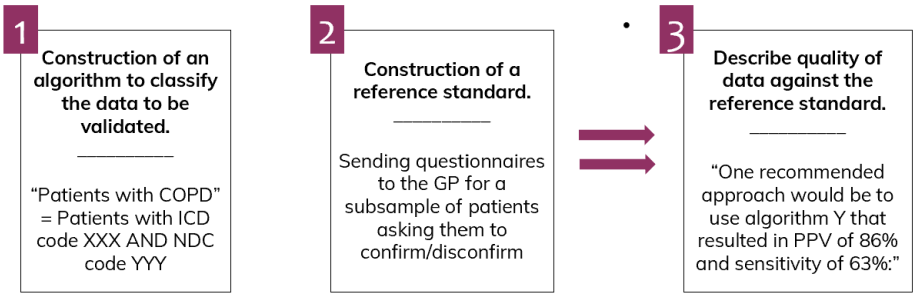

Figure 2: The three steps of data validation against an external reference standard

To support the fitness-for-purpose of the database, the researchers aimed to validate the recordings of 'acute exacerbations of COPD' in the CPRD. The first step was to export from the overall database all and only the events of 'acute exacerbation of COPD'. Often the translation of this medical term into a representation used in a database is not straightforward and can be approached in various ways. At least, it requires combining the medical codes from different medical terminologies, such as the ICD dictionary, and accounting for variability in reporting practices. The FDA guideline refers to these aspects as the 'operational definition'. We might also call it a classification algorithm because it classifies events into acute exacerbation of COPD or its absence using a rule-based decision procedure. Strictly speaking, such the practices validate algorithms rather than data. The question that a validation study seeks to answer is how well different classification algorithms perform in capturing true events, and only true events, within a specific database. In this case, the true event is acute exacerbation of COPD.

Classification algorithms can compensate for misrepresentation in the data. For example, if the classification algorithm not only requires a COPD diagnostic code but also a recurring prescription of treatments that are indicative of COPD,

the extracted data can minimise problems of reporting errors in the diagnosis. By contrast, if an algorithm broadens the scope by requiring only a highly predictive test result as the criterion, one might avoid cases of misdiagnosis at the point of medical judgement. In short, the medical term 'acute exacerbation of COPD' can be operationalised into several different rule-based algorithms for a single database. Not only might the database use different terminologies to represent the same event; these representations might also contain errors, which a more sensitive or specific classification can compensate for. Depending on how broadly or narrowly the classification algorithm is constructed, one risks capturing either more false-positive or false-negative events, respectively. Estimating the number of false negatives and false positives for different algorithms is precisely the purpose of a data validation study. In highly curated databases, the construction of such algorithms could be less complex because data curators already accounted for reporting variability and other issues. As argued in Chapter 4, data curation is limited what it can do about problems of misrepresentation or missing data. Therefore, even in well-curated databases it might be worthwhile to use more complex classification algorithms.

Let's take a close look how the researchers addressed this question in the first example. The medical code systems commonly used in UK general practices are called Read Codes. Based on these codes, the researchers constructed 15 algorithms that plausibly captured acute exacerbations of COPD with different combinations of diagnostic codes, commonly administered treatments and typical symptoms. The team assembled a list of 1043 different medical codes that were used to construct the 15 algorithms. Within these 1043 codes, only two represented the diagnosis of acute exacerbation of COPD. The rest (1041 codes) were either codes for commonly administered treatments for COPD exacerbations (47 codes for corticosteroids, 822 for antibiotics) or represented symptoms such as lower respiratory tract infection (77 codes) and breathlessness (26 codes). Judging the plausibility of these algorithms required expertise from respiratory physicians, primary care physicians with experience in UK primary care, and epidemiologists with experience in the design of large UK primary care database studies. The team then used a previously validated algorithm to preselect only the patients with a COPD diagnosis between 2004 and 2013. They then recorded how many times these patients experienced an acute exacerbation of a COPD event according to the 15 different algorithms. This constituted the first step of the three-step procedure to validate the data.

The second step was the construction of a reference standard, which was the external standard against which the performance of the algorithms was assessed. The standard was assumed to be an accurate representation of the patient's health or disease state. To construct this reference standard, the researchers selected 1600 patients randomly from the previously constructed sample of all patients with acute exacerbation outcomes. Every patient could experience more than one acute exacerbation during the nine-year period of the data. Therefore, the researchers randomly selected up to ten instances, for each patient, that any algorithm classified as an event of acute exacerbation of COPD. The research team then contacted the treating physicians for the patients and asked them to verify the selected events for their patient. The doctors were asked to notify the researchers about any additional event that was not on the list. The researchers additionally asked for any additional source material that supported the physician's judgement, such as test results or hospital discharge letters. This information was reviewed by two respiratory physicians and was used to construct the true classification of exacerbation events for the subsample of patients.

In the third step, the events were categorised into true and false positives or negatives respectively. The performance of each algorithm was calculated in terms of its positive predictive value and its sensitivity (see Table 4). The team counted as true positive cases for an algorithm all events that were a) part of the subsample, b) identified as an event by the algorithm and c) confirmed by the reference standard. False positives met the first two criteria but were not confirmed by the reference standard. This approach allowed the team to calculate the positive predictive value of an algorithm (i.e., the proportion of true events among all the positive events classified by an algorithm). For estimating sensitivity, the team examined how many true cases each algorithm missed. First, it was necessary to learn about the baseline of all true events in the subsample of events. They classified every event as true if it was a) identified as an event by any of the 15 algorithms and b) confirmed by the general practitioner, OR if the event was listed as an additional event by the general practitioner. False negatives could now be defined as the events that were a) not identified by the specific algorithm (but were identified by at least one of the other 14 algorithms); and b) confirmed by the general practitioner OR listed as an additional event by the general practitioner. Because the validated subsample was random, the team could extrapolate the positive predictive value and sensitivity from the subsample to the entire event population statistically.

Table 4: Classification of events into true and false positives or negatives

	GP assessment positive	GP assessment negative	
Classified by specific algorithm	a True positives: identified by the specific algorithm and confirmed by GP	c False positives: Identified by the specific algorithm and not confirmed by the GP	Positive predictive value: a/(a+c)
Not classified by specific algorithm	b False negatives: Identified by <u>any</u> algorithm (except the one being tested) and confirmed by GP, OR listed by GP as additional event.	d True negatives: Not assessed	Negative predictive value: d/ (d+b)
	Sensitivity: a/(a+b)	Specificity: d/(d+c)	

Sensitivity and positive predictive value commonly trade-off against each other. The stricter the criteria in an algorithm to include a case as a true positive, the higher is its positive predictive value because it rarely includes false positives. At the same time, such narrow algorithms will miss many true events, falsely classifying them as false negatives. The results in this case illustrate this. They found that the positive predictive value varied between 61% and 97%. As expected, algorithms that combined symptoms or medical outcomes with the prescription of a medicine had a high positive predictive value, above 90%. Broad criteria, such as the sole use of COPD-related antibiotics, had low positive predictive value.

Sensitivity was generally below 30%, with some algorithms performing as poorly as below 2%. Algorithms with high positive predictive value displayed sensitivity of no more than 25%, with one as low as 1.7%. One algorithm had a notably high sensitivity of 70% but a positive predictive value of only 60%. The team also tried to disjunctively combine all algorithms that showed a positive predictive value above 75%; this approach increased the sensitivity up to 63% while maintaining a reasonably high positive predictive value of 85%. The authors concluded their study with recommending the use of composite algorithms. Additionally, they recommended not using definitions containing COPD-specific medicines unless combined with a relevant medical diagnosis, as such definitions had low predictive value and risked misclassifying events as exacerbations when they were not. Following the authors, such a strategy has been pursued by some studies (Rothnie et al. 2016).

Compared with other validation studies, this study has several valuable features that make it a good validation study. First, the authors went to great lengths to construct different options whose performance could be assessed against each

other. This not only allowed optimising the final strategy by combining different algorithms, but it also allowed estimating the sensitivity of the algorithms and not only the positive predictive value. Secondly, by using a random sampling approach for the subset of events to be validated, the authors could extrapolate the measures to the entire event population using reliable statistical techniques.

However, there is a problem with the reliability of the reference standard. We can imagine that when the general practitioner was asked to confirm or disconfirm the diagnosis, she might simply check whether the diagnosis was recorded in the same EHR system that fed the data into the databases. The reference standard would then no longer be truly external to the data that needs to be validated but introduces circularity into the procedure. Given this circularity, algorithms that include the information that a general practitioner has used to verify the events would have a high positive predictive value, solely because of the circular procedure to validate that data. A second issue of the reference standard is that it cannot avoid various sources of misrepresentation. Most importantly, it cannot exclude that the patient was misdiagnosed by a physician to begin with. Thus, the reference standard used in this case was problematic because of the unreliable construction method and because it might not be a valid representation of the patient's true disease state.

The second example I briefly examine took a slightly different approach, precisely to address the problem of misdiagnosis in the reference standard. The study aimed to validate data recordings of COPD diagnosis within the same database discussed above. The difficulty in identifying true COPD patients from an electronic database is that the diagnosis of COPD is not indicated by a single test but usually requires clinical judgement based on a combination of physical examination and health history. Therefore, the estimated proportion of misdiagnosis is as high as 20% (Quint et al. 2014). The research team aimed to address the problem with the reference standard they used. As in the above study, the researchers sent out questionnaires to general practitioners, asking them to verify the diagnosis and send copies of all relevant source documents – such as test results and letters. In this study, they clearly stated that the gold standard against which they assessed the algorithms was not the clinician's judgement but the review of the source documents by two expert respiratory physicians and their clinical judgement. For the cases where additional material was available, the general practitioner's diagnosis was verified against the expert's judgements. In cases of divergence, the experts corrected the diagnosis in the reference standard to match their own judgement. Such expert judgement differs from the expert judgement

applied in clinical trials, because it occurs asynchronously to data collection and is distant from the patient. One might wonder how experts can judge from afar a patient they have never seen, based on potentially incomplete data. The most serious obstacle to this approach however is a practical problem, namely that source data on most data variables and/or patients is simply not available. In this particular study, among the 696 patients who were classified as COPD patients by any of the algorithms, additional source data were received for only 39% (Quint et al. 2014). For the other 61%, the only information the experts obtained were the physicians' answers to this simple questionnaire:

1. Do you think this patient has COPD? Yes / No / Uncertain
2. What was the diagnosis of COPD based on? (please circle all that apply)
   - smoking history
   - symptoms
   - spirometry
   - other (please describe)
3. Has a respiratory physician seen the patient and confirmed a diagnosis of COPD?
4. Does the patient have any other respiratory condition? If yes, then what?

(Quint et al. 2014, supplementary material).

It seems implausible that even an expert physician could judge a patient's true disease status accurately based solely on information from the questionnaire. Nevertheless, the study included all patients in the reference standard, both those with and those without independent information. The result was a reference standard that was a mixture of expert judgement and (potentially circular) confirmation by general practitioners. The option of including only those cases where source data was available was, however, not viable. They would have risked ending up with a heavily biased estimate, for two reasons: first, misdiagnosis for cases where diagnostic reports are available is less likely. This is evidenced by the high positive predictive value of 95% that the team found when validating the general practitioners' diagnosis against the experts' diagnosis for only the subset of cases where additional data was available. Second, the positive predictive value based on asynchronous expert judgement can be manipulated rather easily. If experts only receive a sample that was already positively classified by the algorithm, every case rated as a true case would increase the positive predictive value. People trying to demonstrate high accuracy of the data could leverage this point. Another limitation of this reference standard is also that it cannot account for misrepresentations that are already contained in the source data, such as poor measurement equipment, poor measurement performance at the site or incomplete or unfaithful reporting by the patient of his or her health history, and many more. Asynchronous expert verification of data against source data is

generally one of the most reliable techniques. However, it suffers from epistemic biases related to sampling procedures, potential bias in expert judgements and any problems of misrepresentation that may be contained in the source data.

Constructing reference standards by sending surveys to clinicians seems a fairly common practice. Yet it is a laborious approach and entails many practical hurdles. Therefore, the FDA explicitly mentions a third approach as an alternative option, namely the use of another routine data source. I therefore complete my analysis with a final example.

Kahn et al. introduced the idea of a *relative gold standard* and illustrated this approach by validating data entries about the ethnicity of patients. They compared the ethnicity classifications of patients in an electronic medical record database with the ethnic classifications of a subset of patients who were also registered in the Hematology, Oncology, Bone Marrow Transplant database (HOB-DB) at a children's hospital in Denver. They extrapolated the results to the entire database based on a similarity assumption (Kahn et al. 2010, p. 357). What is interesting about this example is how the researchers supported the assumption that the HOB-DB could serve as a reference standard: Their judgement was based solely on the author's knowledge about the database. Key characteristics were the data-recording procedures, available staff and training requirements as well as the database's important role for other users (e.g., clinical trial recruiters). These features were deemed indicative of high-quality data. To illustrate, here is a lengthy excerpt from the publication:

> [T]he data contained within the HOB-DB is collected by dedicated research coordinators and data collection personnel with oversight by a full time database coordinator. [...] Because of the critical role of the HOB-DB on multiple departmental missions, substantial efforts are expended to ensure an extremely high level of data quality, including detailed data collection procedures and extensive validation checks [...]. One data element that is deemed especially critical to the HOB investigators is the accurate assessment and recording of patient race [...]. This is a key element for NIH grants and for clinical trials recruitment. Substantial efforts have been put in place to ensure that HOB personnel are trained to assess and record this data element. Thus, race data are highly accurate in the HOB-DB, making the HOB-DB a relative gold standard for patient race. (Kahn et al. 2010, p. 357).

It is knowledge about the database's governance, usages and funding incentives that justifies the database's standing as a relative gold standard for a particular data item of interest. Kahn and colleagues call this 'meta knowledge'; in the EMA's framework, such factors are called the 'foundational determinants' of data quality. The researchers specifically looked at the drivers for quality of the data on ethnicity

and did not rely on a general purpose-agnostic assessment. The reliance on knowledge about the causal drivers of data quality is nevertheless problematic, because it is too unreliable and imprecise to quantify the data's quality into clear estimates regarding its accuracy. Kahn and colleagues were aware of this limitation and duly called the reference standard a 'relative gold standard' – to indicate that the accuracy measures obtained from such standards indicate only the relative improvement over the reference standard and not true accuracy.

The problem is that such relative quality measures cannot justify that data is sufficiently accurate if it is unknown how accurate the reference standard is. At best, relative measures could justify that some data is *in*sufficiently accurate. However, if researchers err about the fact that the reference standard is more accurate than the data that is validated, the relative reference standard cannot even justify that latter. Rather than being a reliable source for data validation, other routine data sources as reference standards only push the question of data quality a step further away. Despite these shortcomings of using one routine data source to validate another, it seems a generally accepted approach. Indeed, the FDA's guidelines proposes using routine data sources as the reference standards, for rather weak reasons:

> For prescribed medications used in outpatient settings, dispensing or billing data would tend to be more accurate than most EHRs in reflecting exposure to a drug by documentation that the prescriptions were filled. In such cases, validation of EHR prescribing data by examining medical claims data may be warranted. For drugs administered in the health care setting (e.g., vaccines, injectables, blood products), administration recorded in the EHR may provide more complete information than is available in medical claims records. In these cases, it may be useful to validate medical claims data by examining the EHR. (US Food and Drug Administration 2021b, p. 17)

Even if 'billing data tends to more accurate than most EHRs', such a relative improvement on quality is insufficient to quantify data quality. These measures do not reliably represent data quality. In the case of the SRTR, which corrected its mortality data in line with the data in the Death Master File, the FDA applied a similar reasoning. Yet, the FDA did not elaborate on which causal drivers of data quality made them deem it trustworthy. Without going into the details here, a quick search on the Social Security Death Master File indicates the following two key quality drivers: first, accuracy of this data is of economic interest, e.g., to prevent identity theft and stop benefit payments from and to deceased people. Second, the Social Security Administration has stringent reporting requirements and

verification processes in place to ensure data quality, including the verification of death certificates and cross-checking information from multiple reporting sources. In the absence of any incentives for the government to forge such data, one might be satisfied with the FDA's assessment that the Death Master File is indeed a 'trusted data source' on mortality data. However, the Death Master File is an exceptionally specialised database that only collects data on mortality, which is a very rare case.

Table 5: Common problems reference standards for data validation

Reference standard	Reliability of the method	Validity of the representation
Clinical judgement using questionnaires	Circularity	Misdiagnosis and measurement errors
Expert judgement using source data verification	Sample biases and conflict of interest, misdiagnosis by asynchronous expert judgement	Measurement errors
Other routine data sources using meta-knowledge	Fallible und insufficient informative knowledge	Misdiagnosis, measurement error, coding errors

To conclude, the three examples use three commonly used reference standards in data validation studies. The first is judgement by the treating physician, the second is verification of source data by expert clinicians and the third is the use of a second (relatively more accurate) routine data source. All come with epistemic problems because they use unreliable methods to construct the reference standards or because they are not good representations of the truth. (For a summary overview, see Table 5.) If the reference standards are unreliable, the quality metrics that are constructed on their basis are also unreliable, and so are the judgements about the data's fitness-for-purpose. Data quality assessments use various techniques and a range of different quality metrics. Some are as simple as checking whether values conform to certain logical constraints (e.g., discharge date is later than admission date), while others are built in comparison to laboriously constructed external reference standards (e.g., the positive predictive value of COPD diagnostic codes relative to expert clinicians' judgements). I have shown that two common strategies also entail common problems and are therefore often unreliable.

This is not to say that all instances of data quality assessments are unreliable; with the right data and skills, a thorough purpose-specific quality assessment might well produce a reliable quality assessment. My epistemic study of some of these

practices has shown how complex such a purpose-specific quality assessment is and how easily data fail to be a reliable reference standard. This work supports the view that we should at least be sceptical about the reliability of fitness-for-purpose claims. The notion of fitness-for-purpose already makes an allowance for reduced reliability of the data; therefore, it seems risky if the methods that establish the reliability of data are themselves unreliable. The expectations about the data's reliability already make an allowance for decreased reliability (we paid the price for power, so to speak). In Chapter 6, I deepen the problem by arguing that data quality assessments cannot be embedded in the network of regulatory oversight precisely because of their local contextuality.

3. A critical study of the power of real-world data

An essential motivation for turning towards a fitness-for-purpose approach is the promised gain in epistemic power. Following Wilholt, a loss of reliability might be justified by a gain in epistemic power understood as the rate at which a method produces results (Wilholt 2016). In the previous section, I have criticised that real-world data fails to meet explicit expectations about their reliability because the methods used to establish the reliability of data are themselves unreliable. Therefore, the failure to meet the expectations are serious. In this section I turn towards the desiderata of power. I established in section 1, in the context of health research, the power of real-world data raises two promises: The first is a hope for accelerating the evidence-generation process to allow patients early access to potentially life-saving medicine. The second is the hope for less expensive evidence generation that would enable redistributing the limited resources to foster more innovation. Intuitively, real-world data seem to be the epitome of a powerful research approach. As Leonelli noted, the mobility of such data is one of their greatest promises (Leonelli 2020). It can travel across users to be used for various purposes. In addition, 'secondary use data' is perceived to be a readily available by-product of routine data collection processes. In this last section, I challenge the idea that real-world data can fulfil these promises.

3.1. Accelerating access to medicines: Who is in a rush?

The first perceived advantage of real-world data is that they accelerate the generation of evidence. Rapidly obtainable evidence is valuable in cases where

prolonged ignorance can cause major harms, and fast access to medicines has potentially life-saving consequences. Stegenga has cast doubt on the assumption that acceleration programmes deliver on their expectations by arguing that the new medicines are barely the life-saving good that people hope for (Stegenga 2017). Others have argued that the EMA's adaptive pathway programme cannot fulfil the expectations about its promised reliability (Davis et al. 2016). It is undoubtedly true that retrospective real-world evidence can be generated rapidly. In the case of the approval of Prograf for lungtransplants, the research team at Astellas reported that it took them only 20 months from planning the study until the FDA's decision. In comparison, it took Astellas 10 years to get Prograf approved for heart transplant recipients through the regular path (PSI RWD SIG 2021). However, according to Astellas, the use of real-world evidence for lung transplants not only accelerated the approval but made it possible in the first place. Tacrolimus for lung transplant recipients was granted orphan designation, with only 2500 patients per year in the entire US, and because of the small population Astellas argued that a randomised experiment would not be feasible. Without the opportunity to use real-world data, Astellas might not even have sought approval with the FDA.

Although Prograf is highly effective, there is something else puzzling about the FDA's and Astellas' reasoning. Tacrolimus has been used in clinical practice for almost two decades; since 2010, it has even been the standard of care to treat lung transplant recipients. Statistics show that between 2010 and 2017, 79% of adults lungtransplant recipients had already been treated with tacrolimus in the US (Erdman et al. 2022). Similarly, the 2018 annual SRTR data report that is referenced in the FDA review states that 85% of all patients have been treated 'off-label' with tacrolimus-based regimens (CDER 2021). Clearly, this is not a case where access to a life-saving medicine is withheld from patients because we need to wait for evidence. Nonetheless, the FDA granted a priority review because there were no treatments approved for this indication, which constitutes an 'unmet medical need' (CDER 2021). Since patients already had wide access to the medicine, it seems there was no real unmet medical need.

This will be the typical scenario for the use of retrospective real-world data in the second regulatory scheme. If we want data to be immediately available, of course it must be data about medicines that are already on the market and prescribed to treat the disease that is subject to approval. Moreover, it is plausible that high-quality real-world data in particular is mostly available for cases where the medicines off-label use is already well-established, be it informally or in official treatment recommendations. For approvals on the basis of retrospective real-world

data, the benefit to patients is typically marginal, because there is no convincing unmet medical need – and without such an unmet medical need, it seems we do not need accelerated evidence. A valid concern against this argument is that patients can run into issues with reimbursement if treatments are prescribed 'off-label'. Approving the medicines thus relieves patients of a financial burden. I acknowledge that this issue is a problem. Nonetheless, problems with reimbursement have far less severe ethical implications than does the situation in which access to medicines is denied. Hence, the social or ethical value that can be gained from using real-world data seems marginal in this case.

More promising attempts to accelerate evidence development and provide access to truly innovative treatments occur through accelerated approval programmes, such as the EMA's Adaptive Pathways launched in 2014. The programme allows the early approval of particularly valuable medicines based on some preliminary evidence, followed by the confirmation or disconfirmation with prospectively collected real-world data. In these cases, patients really do gain faster access to medicines that would not be on the market otherwise. Whether or not this is to the patient's benefit is a different discussion. The epistemic work to accelerate the process, however, is not done by the use of real-world data but by whatever preliminary evidence is deemed sufficient for the initial approval. Real-world data features only after the medicine gains access to the market. Whether the gathering of post-market real-world data is actually faster than running a trial is not yet certain. The EMA's final report on the programme's pilot phase provides a disillusioning conclusion on the use of real-world data (European Medicines Agency 2016). Some scholars doubt that real-world data can play a role in rapidly decreasing uncertainty during the post-marketing phase (Davis et al. 2016) and it is not clear why such prospectively collected real-world data should be more rapidly available than experimental data. One reason might be that the data is easily available in a bigger volume. However, such data also has higher variability which easily consumes the advantage of having a larger volume of data. Hence, it is plausible that collecting a conclusive volume of real-world data might require an equally long or even longer period than for experimental data. In the case of Prograf, the researchers relied on data that was collected over almost 20 years.

Overall, the reason for using real-world data to generate evidence rapidly does not hold up to critical scrutiny. In the first regulatory proposal, retrospective real-world data can be analysed rapidly, but only for medicines that are already on the market. In the second regulatory proposal, the scheme can provide early access to newly marketed medicines. However, the main epistemic work to accelerate the

process is not done by the use of real-world data, and whether such data can uphold to the expectations to be more rapidly available than experimental data remains to be seen.

3.2.Saving resources: Cheap for whom?

The second perceived advantage of real-world data is that it helps to prevent opportunity costs, because such data are a comparatively cheap by-product of healthcare. However, there are many costs involved in the production of such data that need to be accounted for. Thanks to the pioneering work of Leonelli, philosophers have learned that it is a laborious, skilful and very costly task to render data reusable. It requires work such as structuring the data, cleaning data, coding data with semantic standards or annotating the data with metadata. Building data curation processes that transform clinical data from a 'raw' form, as mostly unstructured text, into so-called FAIR data – which stands for Findable, Accessible, Interoperable, Reproducible – takes years of skilled labour (Leonelli 2016). Countries across the world are investing large amounts of money into initiatives to perform such work. These initiatives include large data-driven projects such as the Beyond 1 Million Genome Project, the All of Us project, EU Darwin, Findata, the European Health Data Space and the Swiss Personalised Health Network, to name a few. Yet the work of such initiatives goes well beyond the data curation itself. It begins with negotiating semantic and technical standards, building secure IT environments for sharing sensitive health data and negotiating legal frameworks and data governance processes. With an increased need for high-quality data, such infrastructures are increasingly also confronted with the need to invest in monitoring or certification (Bernal-Delgado et al. 2022; Daniel et al. 2018). Not only is the development and maintenance of such infrastructures costly, it is paid for largely by public money. A cross-European survey on data quality practices among health data–sharing networks identified 31 initiatives in Europe that had a primary focus on sharing health data. The study also found that almost all of these initiatives were exclusively paid for by public funds (Bernal-Delgado et al. 2022; Daniel et al. 2018).

The direct cost of data infrastructure, however, is just the most visible of all the costs that are attached to the production of real-world data. To decrease the continuous costs incurred by data curation, governments are additionally investing into streamlining data-collection processes at the source. A well-known investment programme is the Medicare and Medicaid Electronic Health Record Incentive Program in the US, also known as Meaningful Use, which was launched in 2009.

The programme allocated USD 34 billion for incentive payments for healthcare facilities to stimulate the adoption of EHR technology in health and the further demonstration of its 'meaningful use', including, among other requirements, the provision of structured data for secondary use. In 2018, the programme was transformed into a revised version of the incentive payment programme, the Promoting Interoperability Programme. As a consequence of the political desire for high quality data 'at the source', healthcare professionals became an indispensable part of the data collection pipeline. Inverso et al. estimated that the costs for clinics to meet the meaningful use requirements amount to USD 184 per patient, with a total of roughly USD 3700 per day per clinic (Inverso et al. 2016). Based on interviews with healthcare providers in Denmark, Green and colleagues showed that the political desire for high-quality reusable health data imposes all kind of invisible 'data work' and invisible costs within healthcare facilities (Green et al. 2022). For example, healthcare professionals must spend more time on multiple testing, collecting data or validating data, which drains considerable resources from patient care (Green et al. 2022). Hence, repurposable health data does not imply merely reusing data from clinics; instead, new data-collection requirements are imposed that follow the needs of the secondary user. Ironically, Green et al. show that the additionally collected data is often not only irrelevant for clinical purposes but even risks harming clinical care through problems such as information overload in the patient's documentation. Efforts to fundamentally reorganise data collection at the source clearly challenge the idea that real-world data is 'secondary use' data, when much of this data never even had a primary user.

Another form of invisible costs occurs at the side of regulatory bodies. Historians and social scientists have argued that the biomedical sciences are governed by a unique complexity and number of scientific conventions – which are indispensable for the good functioning of the community (Cambrosio et al. 2006, 2009; Hauray 2017). With the evolving regulatory standards, the scientific community together with regulatory bodies around the world are working at full speed to develop new standards and accumulate new experience regarding how such standards can be put into practice. The FDA alone has produced eight new guidelines since the launch in 2018. Various initiatives have been funded to conduct experimental pilot projects to accumulate new experience with the type of research that underpins these guidelines. The Get-Real consortium, the RCT DUPLICATE initiative (Franklin et al. 2020b; Franklin et al. 2021) and the EMA's pilot projects to explore the potential of adaptive pathways (European Medicines Agency 2016) are a few examples. That these initiatives involve substantial costs is

acknowledge by Rawlins, a critic of the costly regulatory standards for the industry. It is because of these costs that 'the international community should 'embark on collaborative methodological research', 'because ... [i]t would be perverse to expect the pharmaceutical industry to undertake such a programme: it would only, at least in the short term, further increase the costs of drug development' (Rawlins 2004, p. 363). However, the substantial costs do not disappear just because they are redistributed towards public parties.

There are yet other costs to be expected. Andreoletti and Teira used the legal distinction between rules and standards to draw philosophers' attention to the economic costs involved in adopting a standard-based regulatory paradigm (Andreoletti and Teira 2019). Unlike rules that apply regardless of their justification, the applicability of standards needs to be continuously re-evaluated in light of their justification, which increases the costs of enforcing such standards on a daily basis. I think from discussions in section 2 it should be clear that quality assessment of real-world data cannot amount to rules that apply regardless of their justification. As an aside, Andreoletti and Teira also note that the unpredictability of standards (as opposed to rules) might ironically even increase the costs on the side of the industry, because companies can no longer anticipate what the regulatory requirements entail. Regulatory agencies are sometimes criticised for their high approval rates. Yet one reason for the high approval rate lies precisely in the predictability of their requirements, which directs companies to develop medicines that have a strong chance of acceptance.

A final point is worth making regarding the costs of the EMA's adaptive pathway programme. Within the programme it is foreseen that companies receive scientific advice from the EMA and health technology assessment bodies that are in charge of evaluating reimbursement conditions for products. The main purpose of such advice is to increase the chance that medicines will be directly reimbursed at the time they receive the first market authorisation (Davis et al. 2016). Hence, unlike compassionate use programmes – where companies typically provide early access to medicines at their own cost – within the new adaptive pathway programme these costs are covered by insurance or patients. These programmes would relieve companies from the burden of paying for clinical trials and would support evidence production by redistributing much of the data collection cost to public parties. Additionally, companies would be paid by insurers for the continuous testing of their treatments.

As we can see, throughout the data production and review cycle, real-world data entail considerable costs. These include costs for the negotiation of data

standards or the setting up of data governance processes; costs for the laborious collection of data in the clinics, and for the skilled curation and validation of data in data warehouses; and costs for the negotiation of new regulatory guidelines, their empirical testing and the continuous reinterpretation of such standards to enforce them. Although real-world data promises to be 'secondary use data', it appears that much of this data in reality never had a primary user. Rather, public parties are producing new data according to the needs of secondary users. It appears that the costs are not reduced but largely redistributed towards public parties.

The common view is that most of these costs are an investment into the future and will decline over time. This is precisely the seductive aspect of this idea: Once data has been made reusable, it can be reused for all kinds of different purposes, and in the long run our investments will pay off. Yet the facts here remain unclear. Unless real-world data gains considerably in quality, we have reason to suspect that the number of purposes for which a particular set of real-world data is fit-for-purpose will be minimal. Yet bringing the data up to a high level of quality and maintaining it is certain to be costly. To underline my point, I return to our case study on the approval of Prograf. The data submitted to support the medicine's approval was deemed fit-for-purpose, and the submission was accepted as 'adequate and well-controlled', in line with the FDA's requirements. Yet the data had various problems, two of which I have mentioned in this chapter, namely the failed primary analysis and the lack of data on dosage. In Chapter 4, I presented additional issues. The reason why the FDA nevertheless accepted the evidence as adequate and well-controlled was that they could rely on various contextual evidence, such as the fact that lung transplants without therapy have a high mortality rate and Prograf is effective at preventing mortality. Moreover, the FDA relied on the mechanistic comparability of earlier interventions to fill relevant evidence gaps, such as missing data on dosage (for a detailed discussion see Chapter 4). The issue here is not the reliability of the overall evidence. The issue is that the contextual evidence did most of the epistemic work, but in most other cases, such contextual evidence will be unavailable. For most diseases, the natural course of disease is much more complex and heterogeneous than the natural history of lung transplant. In addition, most medicines today are minimally effective, as exemplified by a list of widely used medicines with questionable effectiveness in line in Stegenga's medical nihilism (Stegenga 2018). Moreover, the mechanistic comparability between indications was another exception. Hence, data from the SRTR was only fit-for-purpose in light of already conclusive and rare contextual circumstances.

Nonetheless, overall, the SRTR is a promising data source. It is a well-established infrastructure that has been managed by experts for almost 40 years. Moreover, establishing the effectiveness of Prograf for the general population is probably one of the most basic usages for this data. If data from this database is only fit-for-purpose in light of rare contextual evidence, my concern is that we would widely overestimate the number of purposes that the *same set of data* can actually fulfil. These concerns are supported by two other specialised studies for which the SRTR failed to provide fit-for-purpose data. First, Yanik et al. directly evaluated the accuracy and completeness of the SRTR data on the reporting of cancers in transplant recipients. Cancer reporting is mandatory in the SRTR database because transplant recipients have an increased risk of cancer due to the lifelong use of immunosuppressive therapy. However, the researchers found that only 36% of cases reported in other cancer-specific registries were also reported in the SRTR database (Yanik et al. 2016). Second, Sawinski et al. used the SRTR database to examine the effect of protease inhibitor (PI)-based regimens to treat HIV by studying the outcomes among transplant patients; they linked the SRTR database with pharmacy disposal data. The results failed to confirm an effect of PI-based regimens on the risk of acute rejection, although such a finding was clearly expected (Sawinski et al. 2017).

Perhaps the problem is that we are not thinking far enough into the future. Eichler and colleagues argue that the 21st century will see a shift from last century's 'blockbuster drugs', such as statins – which are mostly chemicals with small effects for large populations – towards more complex medicines. Hence, we might soon witness a shift towards 21st-century medicines such as biologicals, cell or gene engineering therapies and drug–device combinations. The trends are likely towards ever smaller target populations in rare diseases and in biomarker-driven drug development in oncology. These trends are already visible in the type of approvals since 2000 and are supported by horizon scanning of regulators (Eichler et al. 2021). It is in this future where real-world data will no longer be optional but one of the only ways to generate evidence about populations that are becoming too small – and treatments that are becoming too complex for randomised trials.

If Eichler and colleagues are right in their forecast, these trends would change the way evidence is generated to support the effectiveness of treatments. In such a future, the main use of real-world data would be as an external control arm for small single-arm trials. Empirical research indicates that this is already common practice in the production of evidence for cancer treatments with orphan designation (Gloy et al. 2023). In this future scenario, the use of real-world data is

based on the premise that its use will become a necessity in instances where randomised trials are unfeasible. In those cases, evidence from real-world data is a priori more reliable and powerful than having no evidence at all. However, in Eichler's future vision, these 21st-century medicines will be embedded in a shift towards the further specialisation of medical practices. Such a shift itself offers the opportunity to generate high-quality real-world data at the source:

> [I]n the future, specialized tertiary care facilities should be expected and held accountable to implement a high level of patient documentation that enables generation of high-quality real-world data, and ultimately the development of a 'learning health care system' with the ability to provide increasingly robust assessments of drug effects over time. (Eichler et al. 2021, p. 1214)

I think it should be clear by now that such data generation envisioned by Eichler et al. will entail considerable costs for patients and public healthcare providers.

The arguments in this section support the critical view that the use of real-world data might not reduce the costs but rather distribute them towards public third parties, including governments, regulators, healthcare facilities and patients. Current problems with the use of such data suggests that the power of this data is considerably overestimated. Even a well-established registry barely supported a simple use of its data. Making data reusable for more purposes requires substantial investments into collecting more data, better structured data and different data. Ironically, we are caught in a cycle in which our attempts to increase the power of data actually decrease that power.

Chapter 6
The trustworthiness of real-world data

The 'Surgisphere scandal' stands out as a troubling instance of fraudulent research conduct during the early days of the COVID-19 pandemic. The private company Surgisphere claimed to possess evidence about the effectiveness of hydroxychloroquine to treat COVID-19 from EHR data from almost 100,000 patients across 671 hospitals on six continents (Mehra et al. 2020). It did not take long until suspicions about the authenticity of the publication by Mehra and colleagues piled up, which prompted its retraction. Yet, in the midst of an ongoing pandemic, the publication still had a major impact on public health policy (Offord 2020). Among the numerous flaws in the publication that aroused suspicion regarding its authenticity were the nearly impossible scale and the claimed sophistication of the database. An illustrative excerpt from their publication describes the infrastructure as follows:

> The Surgical Outcomes Collaborative [...] ensures compliance with the US Food and Drug Administration (FDA) guidance on real-world evidence. Real-world data are collected through automated data transfers that *capture 100%* of the data from each healthcare entity at *regular, predetermined intervals*, thus reducing the impact of selection bias and missing values, and ensuring that the data are current, reliable, and relevant. *Verifiable source documentation* for the elements include electronic inpatient and outpatient medical records. [...] Collection of a *100% sample from each health-care entity is validated against financial records and external databases* to minimise selection bias. *[...] Data have been collected from a variety of* urban and rural hospitals, academic or community hospitals, and for-profit and non-profit hospitals. (Mehra et al. 2020, p. 2, my emphasis)

In the preceding chapters, I delved into the epistemic risks of claims about data quality and illustrated the risks involved and the extensive effort needed to substantiate such claims. Given that discussion, the sheer implausibility of maintaining such a sophisticated database should be self-evident. The logistical and legal obstacles that would need to be surmounted to assert possession of verifiable source documentation on such a vast scale are considerable. It would require extensive effort to verify a sample from each healthcare entity against an external database. In the Surgisphere scandal, it seems that Mehra and colleagues not only

fabricated data or claims about the data's quality but falsified the entire database itself (Offord 2020).

In light of such a disturbing scandal, real-world data raises major concerns in the clinical research community regarding the risk of fraud. Among philosophers of science, John argued explicitly against the use of real-world evidence for health policy making on the ground that such evidence is not 'robust' against manipulation (John 2021). Despite its immediate plausibility, the challenge such an argument must address is to show why real-world evidence should be more susceptible to this concern than data from clinical trials. Philosophers have unravelled various problems with clinical trials, and the problem of data manipulation is one of them (Borgerson 2009; Stegenga 2018).

In 2017, the editor of the journal *Anaesthesia*, John B. Carlisle, became so concerned about the potential prevalence of untrustworthy data in clinical trials that he began scrutinising all clinical trials submitted to the journal. He used the label 'zombie trial' to categorise a trial 'when the extent of data fabrication lost my trust'. Specifically, he lost trust in studies if he thought that 'the authors had compromised science by lying or being incompetent' (Carlisle 2021, p. 477). Carlisle analysed 153 trials at the level of individual patient data between 2017 and 2020, of which 26% made it into this category. Interestingly, among the 373 trials that he analysed at the level of published summary statistics, he identified only 1% of trials as zombie trials. This finding indicates that many questionable studies remain undetected unless someone meticulously scrutinises the data.

These results have spurred a new discussion about the prevalence of fraudulent data in clinical trials going undetected. In response to these concerns, several trustworthiness screening tools were developed for clinical trials to identify trials with untrustworthy data in a more systematic manner. Trustworthiness screening tools extend the more commonly known risk-of-bias assessment of clinical trials by assessing factors such as good research governance, plausibility of baseline characteristics, plausibility of the study's feasibility and the plausibility of results.[21] The first applications of such tools to a systematic selection of trials found that about 25% failed the check (Alfirevic 2023).

The challenge I tackle in this last chapter is how we could be confident that claims about data quality are made with sufficient competency, are reported

[21] Two of the better known tools are the Cochrane Pregnancy and Childbirth Trustworthiness Screening Tool, developed by said Cochrane group, and the REAPRAISED checklist developed by Grey et al. Grey et al. (2020). For a discussion of risk-of-bias assessments, see Chapter 2.

sincerely and do not involve controversial value judgements. Hence, I attempt examine how such claims might be trustworthy despite the community's susceptibility to conflicts of interest. I first develop a proposal regarding how clinical trials have (successfully) responded to this issue. Then I argue that the same solution cannot be applied to solve the problem of trustworthiness of real-world data studies. My proposal about the trustworthiness of clinical trials is only a rough sketch of what should be a much more rigorous discussion of various conceptual issues involved in trust in science or trust in institutions, and it is beyond the scope of this chapter to address any of them. Yet, I hope to draw philosophers' attention to the fundamental function that regulatory institutions, professional roles (such as monitors) and tools (such as audit trails) play in ensuring the trustworthiness of experimental data in clinical trials and the proper functioning of the clinical research community.

I first develop my proposal on the trustworthiness of clinical trials by closely examining once more the crucial guidelines of GCP by the ICH. From that examination, a social epistemology perspective emerges, in which trust in clinical trials is grounded in the mandatory involvement of impartial experts throughout the research process. These impartial experts exhibit roles within institutions – such as ethics committees, independent monitors or inspection authorities – that use tools such as monitoring reports, audit trails or trial protocols to execute their role. Building on John's two-premise model of trust in science, I propose to understand these instruments and roles as signs of a 'well-ordered community'. I then develop three arguments to show that trust in real-world data cannot rest on the same grounds. The first argument hinges on the contingent fact that currently none of these roles or tools are commonly employed to produce real-world data. The second argument rests on the normative problem that employing these tools and agents to augment the data's trustworthiness comes at a high and potentially undesirable price: an increase in trustworthiness is a trade-off with the data's main epistemic advantage, its epistemic power. The third and main argument is a conceptual argument building on the methodological features of data validation. Through another close look at data validation practices, I highlight its deep local contextuality and value-ladenness. I develop the view that the local contextuality and value-ladeness of these practices are substantial obstacles for embedding the production of claims about data quality into the larger well-ordered clinical research community. Overall, I suggest that the ongoing evolution of evidence standards seems to be a rather radical shift from the paradigm of regulatory oversight of rule-based execution of research to a paradigm of trust in local experts.

This shift suggests that the clinical research community might bring about new foundations of trust and different ways of coordinating its cooperation.

Section 1.1 develops the proposal for the trustworthiness of clinical trials. Section 1.2 presents the first two arguments against the trustworthiness of real-world data. Section 2 turns toward data validation. I first elaborate on the local contextuality and deep value-ladenness of the method (section 2.1) and then develop the third argument against the trustworthiness of real-world data (section 2.2).

1. Trustworthiness of clinical trials

Trust in general and trust in science in particular have been popular topics in the philosophy of science for some time. In this literature, philosophers are commonly interested in what trust is and how it is distinct from related concepts, such as reliance. They also consider whether trust relationships occur only between individuals or can also occur between individuals and institutions or between individuals and communities. It is not the project of this chapter to tackle conceptual puzzles about the notion of trust or trustworthiness. What I am interested in are the reasons we could possibly have to believe that data in clinical research has been gathered with sufficient expertise and reported sincerely and has not been compromised by inacceptable value judgements. I consider such reasons a necessary requirement for the (justified) uptake of evidence by various stakeholders. My goal for this section is to provide an analysis of the instruments, actors and rules employed within the clinical research community that constitute such a warrant. Who or what exactly the bearer of trustworthiness is, or whether the subsequent relationship is one of trust or mere reliance, are questions that are irrelevant to that analysis.

I begin with the three common sources of mistrust alluded to above: concerns about expertise, sincerity and acceptable value judgements. In the context of science, we are mostly interested in a special kind of trust called 'epistemic trust', meaning that we trust a person as an information provider (Wilholt 2013). That is, trust as a three-place predicate understood as trust in someone to do something. As a basic requirement, if we place epistemic trust in someone, we need to be confident that the person who generated the knowledge did so with the necessary competence and reported the results sincerely. A violation of either of these two requirements was what motivated Carlisle, the editor of *Anaesthesia*, to classify some clinical trials as zombie trials. That is, in such cases he felt that 'the authors

had compromised science by lying or being incompetent' (Carlisle 2021). Most philosophers think of epistemic trust as having a performative component (competence) and a normative component (sincerity).

In recent decades, philosophers of science have made a convincing case that the normative component of epistemic trust is more nuanced than just the absence of outright fraud. Philosophers widely acknowledge that ethical, political or social values cannot be eliminated from the generation of scientific knowledge. A forceful argument for this position, the argument from inductive risks, departs from the insight that no amount of scientific evidence can ever be fully conclusive about the truth of a scientific hypothesis but always remains uncertain to a degree. Thus, researchers must determine what is considered an appropriate level of evidence to accept a hypothesis as true or reject it as false. Settling on evidence levels, however, involves trade-offs between different inductive risks. The most prominent example is the risk of accepting a hypothesis as true when it is in fact false (the risk of false positives) and the risk of rejecting a hypothesis as false when it is in fact true (the risk of false negatives). The question then becomes how scientists balance – or ought to balance – these different risks. A common answer is that they consider – or ought to consider – political, ethical, or social goods involved when any of these errors are made. The argument was first proposed by Rudner, focusing on the problem at the stage of hypothesis testing. More recently, the debate has been revived and strengthened by Douglas (2000; 2009), who argued that researchers encounter similar trade-offs throughout the research process.

The implications of these arguments for the question of trust and trustworthiness are twofold. First, it has become a consensus among philosophers of science that the ideal that science is value-free is unattainable and perhaps even undesirable (Douglas 2009). Consequently, the absence of values in scientific research cannot be the basis for any claims about the objectivity and trustworthiness of science. Rather, it is 'the right use' of values that demarcates good science from bad science.

Second, these arguments have added another source of mistrust, namely uncertainty about which values scientists emphasise to make the various judgements in their research. Some researchers who stress values such as the wellbeing of patients might deserve our trust, because those values are aligned with our own. Researchers who emphasise other values – such as the profitability of products – do not deserve our trust. Irzik and Kurtulmus accordingly distinguish between basic and enhanced epistemic trust to capture the difference between trust

that only considers expertise and sincerity versus the enhanced version that also considers the alignment of value judgements (Irzik and Kurtulmus 2019).

Since trustworthiness involves the right moral attitudes of individual scientists, strategies to foster trust in science have often focused on the virtues of individuals, such as honesty, sincerity or epistemic self-assessment. Measures such as transparency and public involvement have attracted considerable support among philosophers as promoters of trustworthiness of researchers. Irzik and Kurtulmus, for example, discuss hybrid deliberation forums – such as consensus conferences, citizens' juries and panels – as indicators of the trustworthiness of research (Irzik and Kurtulmus 2019). Solomon critically assessed the widely hold ideal that consensus conferences in medicine can promote democracy and objectivity (Solomon 2015, chapter 4). Accounts that focus on the trustworthiness of individual scientists, however, fall short of accounting for the difficulties and complex realities of clinical research. In a field that is prone to conflicts of interests, we have good reason not to believe that value judgements are aligned among researchers, authorities and public stakeholders. We might not even have a good reason to believe that scientists have noble intentions to report evidence truthfully. Consequently, scandals in biomedicine have not just resulted in the loss of public trust in specific individual scientists but have also prompted an ever-tighter regulatory oversight and narrow net of globally binding conventions regarding the epistemic and ethical standards that hold researchers accountable. Hence, any discussion about the trustworthiness of clinical research(ers) needs to consider the socio-economic reality in which the research occurs.

Two philosophers who pay justice to such sociological considerations are Stephen John and David Teira. John has developed a two-premise model of epistemic trust in science that considers not only the epistemic but also the sociological conditions of research (John 2021, 2018). The idea of his model is that trust must fulfil a sociological and an epistemological requirement. The sociological premise is that communities that produce scientific claims must be well-ordered, such that the best explanation for a scientific claim meets a set of epistemic standards set by the community. The epistemological premise is that if a claim meets the epistemic standards, I should accept that claim as well. The epistemological premise provides John with the grounds to argue that the scientific community should adopt epistemic standards that are 'broadly acceptable' to maximise public agreement about which scientific claims are true (John 2021, p. 5). In his view, high epistemic standards can play this role of being broadly acceptable. The reason is that disagreeing parties can generally agree that claims

that meet the highest standards are established, whereas parties might disagree about whether claims that adhere to lower standards are sufficiently well established. Assuming that randomised trials are a higher standard than real-world evidence, the political need for maximising agreement about which scientific claims hold gives John a political reason to prefer randomised trials over real-world evidence. That is, we can achieve agreement about the claims that are established with RCTs, whereas parties might disagree about claims that adhere only to real-world evidence standards.

The sociological premise adds another perspective to these epistemic debates. The premise requires not only that the community agrees on acceptable standards, but also that it is well-ordered, in the sense that it adheres to these standards when producing scientific claims. Hence, the sociological premise can bypass all disagreement about epistemic standards: if the community is not well-ordered, we should not accept results, because they might not be the result of meeting some standards, regardless of what these standards are. Hence, we not only need generally acceptable standards but we also need markers of 'well-ordered' communities to foster trust in science. One way to understand John's proposal is that in a community that is not well-ordered, factors such as incompetence, insincerity, conspiracies or misaligned value judgements could be equally good explanations for the community to make certain scientific claims. In our case, the highly commercialised nature of the clinical research community is a good reason to believe that the community is not well-ordered. Hence, insincerity and economic interests are equally good explanations for scientists to assert a certain claim.

Using his two-premise model of epistemic trust, John identifies the real-world evidence standard as untrustworthy based on two premises. The first states that the clinical research community is not well-ordered; the second premise states that real-world evidence is not robust against manipulation by this community. He argues:

> [R]eliance on RWE is problematic because it is very easy for the pharmaceutical industry to 'game' real-world trials (Davis et al., 2016). For example, in real world trials, researchers have a huge amount of leeway in choosing which evidence to include or exclude. In the case of CDF2 [where the use of real-world evidence is allowed] specifically, researchers can present evidence from cancer registries in favour of their claims; however, there are concerns that this data is partial, incomplete, and easy prey for cherry-picking ... [T]he use of such evidence is not 'robust', in the sense that it can easily be manipulated by interested parties (Holman & Geislar, 2018). We have good reason to be wary of CDF2. The epistemic community is not 'sociologically well-ordered'. (John 2021, p. 6)

I share John's concern that real-world data in clinical research poses this problem of trustworthiness. One might wonder, however, why the absence of a well-ordered community poses problems only for the production of real-world evidence and not for the conduct of clinical trials. After all, it is the same community that produces both types of evidence. Teira's work on the impartiality of the randomised method helps to explain this point. He thoroughly analysed the origins of the randomised clinical trial within the regulatory context and argues that the method emerged as an acceptable standard because it ensures the best possible impartiality (Teira 2020). For example, randomisation spares the community from the task of agreeing on the relevant confounders, which fosters acceptability of the method and its results (Martinez and Teira 2021). Blinding as a methodological control prohibits various parties from manipulating the results according to their interests (for an extended discussion see Chapter 1). Hence, in Teira's view, the randomised trial became the preferred gold standard in the community precisely because it is a highly impartial method on which all conflicting parties could agree (Teira 2016). Hence, the adoption of the randomised trial as the evidential gold standard could be a sign of order in an otherwise not well-ordered community, because the method is robust against manipulation.

I want to add a second answer to this puzzle that refers to the social network in which the method is embedded. This network plays a role in protecting data against manipulation, controversial value judgements and even incompetence. This network includes institutions such as ethics committees and regulatory agencies; professional roles, such as monitors or auditors; and various instruments, such as audit trails, monitoring reports and research protocols. I suggest that the relevant clinical research community consists not only of researchers within pharmaceutical companies but also this larger network of institutions and actors that jointly contribute to the good ordering of the community – and hence the trustworthiness of evidence produced by this community. However, as I argue later, any attempt of stakeholders to embed practices of quality assessments for real-world data within this broadened and well-ordered community has to face potentially unsurmountable obstacles.

1.1. Ensuring trust with Good Clinical Practice conventions

The ICH-GCP-E6(R2) is a pivotal document that sets forth globally accepted standards and principles for conducting clinical studies, first published in 1997. It governs the quality of data by creating accountability, defining mandatory

instruments and rules for data collection, storage and reporting. In Chapter 4, I introduced parts of these guidelines in detail and argued that if functions as a gold standard for data quality in the community. The fact that real-world evidence is generated from data that was not governed by these guidelines is, for many stakeholders, evidently a greater source of concern than the fact that such data was analysed using an observational or pragmatic method. During informal conversations with practitioners, I repeatedly heard spontaneous comments such as 'without these guidelines, I have to do all the work by myself', or 'having these guidelines is crucial, and without them we would descend into anarchy'.

Historians and sociologists have emphasised the unique and unprecedented role played by the vast number of conventions in biomedicine. Cambrosio follows Daston and Galison in historicising the concept of objectivity and argues that the volume of regulation (or conventions) in biomedicine has created a new type of objectivity, which he calls 'regulatory objectivity' (Cambrosio et al. 2006, 2009). Hauray shows how the narrow net of hundreds of national, regional and international guidelines – for every type of study and all medical fields – serves to disguise the messy regulatory decision-making process as scientific and objective (Hauray 2017).

Naturally, transitioning to new conventions deeply disrupts the coordination within the community. The transition deprives the community of decades of accumulated experience of working with the previous guidelines. This fact has also been observed by practitioners themselves (Franklin et al. 2019). The current rush in the real-world evidence movement to produce new guidelines, as well as efforts to study and accumulate past experiences using real-world data, illustrate this disruption and the community's need to uphold a 'regulatory objectivity'. In that regard, the historical observation is spot on: The FDA alone has published 9 different guidance documents since the launch of their real-world evidence programme in 2018 (US Food and Drug Administration 2023). Many articles discussing past experiences with real-world evidence in regulatory decision-making have been written (Flynn et al. 2022; Hatswell et al. 2016; Jonker et al. 2022; Mahendraratnam et al. 2022). Initial experiences, such as the FDA's approval of Prograf and the Salford Lung Study, are discussed at length and disseminated as paradigm cases in the community (PSI RWD SIG 2021). The socio-epistemic perspective on conventions in biomedicine holds that it is only a matter of time until the new 'regulatory objectivity' is established, and stakeholders will perceive real-world data as objective as data from clinical trials. New accumulated experiences will serve as guidance in the process of turning such data into decisions.

What the socio-epistemic perspective cannot account for, however, is the epistemic value contained within a particular set of conventions. That is, it cannot explain why the adoption of some set of conventions rather than another might be epistemically preferable. In Chapter 4, I argued that the guidelines for GCP are a good convention to ensure data quality because they increase the likelihood of reliable data with prospective planning and corrective feedback loops. Here, I am interested to see how the norms within these guidelines safeguard the data's trustworthiness by protecting data from manipulation, controversial value judgements and incompetence. I show that the GCP guidelines not only define the epistemic standards for good data but also ensure that everyone plays by the rules and adheres to these standards. The GCP guidelines respond to both – the epistemic and sociological premise – and are therefore an exceptionally pivotal set of rules within the community. Practitioners who are worried about the divergence of real-world data from these guidelines are not only worried about the transition to any new standard, but they are worried because they might loose a source of trustworthiness of data that is difficult to restore.

The GCP guidelines enjoy a high level of legitimacy. They are published by one of the most important bodies to define globally accepted standards in the field of pharmaceutical research and development, namely the International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH). Its primary mission is to harmonise drug development globally in order to facilitate decision-making by regulators and avoid unnecessary duplication. The ICH was founded in 1990 as an informal international partnership between industry and regulators. Today, it is a legal entity, a non-profit organisation established under Swiss law. To strengthen its independence and trustworthiness, the organisation underwent a reform in 2015 to ensure its financial independence from the pharmaceutical industry. Other aims of the reform were to increase transparency in its decision-making process and clarify its governance roles. The development and implementation process of each guideline, building on expert consensus, public consultation and international endorsement, is essential for the legitimacy and acceptability of these conventions. Each guideline is developed by a working group of delegated international experts appointed by the ICH member organisations and obliged to represent the view of their respective organisation. The development of each guideline follows a transparent and standardised five-step procedure. This includes the consensus-building among the experts in the working group about a draft document, further endorsement of this document by the ICH assembly and regulatory topic leaders and a public consultation process

that is open to all interested stakeholders. As part of the implementation, the guidelines for GCP are often referenced in the national laws, making the guidelines legally binding for all clinical trials (International Council for Harmonisation 2021a). The legitimacy of this organisation and its processes contributes to the high standing of its guidelines in the community and reinforces their acceptance.

The main epistemic value of their content is as follows. The GCP guidelines protect data from untrustworthy human behaviour by embedding clinical trials into a larger socio-epistemic context, which includes the mandatory involvement of *impartial experts* throughout the research process. *Ethics committees* approve the research plan before the experiment begins, *monitors* verify that research staff are adhering to the rules throughout the trial and authorities can inspect the trial at any time. Research protocol, monitoring reports, audit trails and training logs are only a few of the essential instruments that the community employs to execute their roles. Moreover, by clearly defining roles and responsibilities within the research team these conventions create accountability for the integrity of the data. All these agents and their roles and responsibilities as well as the instruments they work with are binding rules required by the pivotal conventions of GCP. In these guidelines we can find specific measures that address all three potential sources of mistrust. The next paragraphs elaborate on some of these requirements in detail. All references refer to Revision 2 of the GCP guidelines published by the ICH in 2016 (International Council for Harmonisation 2016).

To ensure expertise, the GCP guidelines have several sections devoted to training requirements, beginning from a high-level principle stating 'Each individual involved in conducting a trial should be qualified by education, training, and experience to perform his or her respective task(s).' (International Council for Harmonisation 2016, principle 2.8). Throughout the document, the area of expertise required for different tasks is specified, including training in GCP, medical training, familiarity with the investigational product and specific expertise for roles such as monitors and auditors who control the ongoing activities. To ensure the research team is knowledgeable about project-specific procedures, handling of electronic systems and other activities that are relevant for data collection, each research member must undergo training. The guidelines further require that CVs and training activities must be documented and reviewed by monitors and, to an extent, ethics committee members.

To address the risk of fraud, the GCP require so-called 'audit trails' for data in regular clinical trials. Audit trails are a documentation of all changes made to the data, including that all changes to the data are dated and initialled by the person

who made the changes. This instrument is crucial as it ensures that data changes could be inspected by an impartial expert in the case of suspicious activities. Another well-known instrument that addresses the risk of fraud is the requirement for a research protocol, where researchers have to plan how data will be collected and analysed (International Council for Harmonisation 2016, principle 2.5). Publication of the research plan ahead of the study mitigates fraudulent techniques, such as the selective reporting of data or the manipulation of significance thresholds with 'p-hacking'. All members in possession of the protocol can verify that data was reported as planned.

The research protocol also helps to constrain the possibility of controversial value judgements in a twofold way. First, once the parties have agreed upon a protocol, all sites are required to conduct the data collection in accordance with this plan, which narrows the range of acceptable value judgements of individual researchers during the conduct of the trial. Second, trial protocols must be approved by an independent ethics committee before the experiment is conducted. Their role is to ensure that the expected benefits outweigh the expected risks and that the conduct is in line with ethical and social requirements. Hence, although the planning of a trial involves various value judgements, the planned experiment has to be reviewed and deemed acceptable by an impartial ethics committee.

Finally, so-called monitors play an essential role throughout the research process. Their responsibility is usually to visit each research site; they not only verify the quality of the data (as discussed in Chapter 4) but also ensure that everyone adheres to all the other rules outlined in the GCP and the research protocol. They review documents (such as training documentation or CVs) and check whether the on-site storage and supply of the investigational medicine is in accordance with the protocol, and they verify the reported data against the available source data on-site (International Council for Harmonisation 2016, section 5.18).

Clinical trials that are conducted according to the rules of GCP thus involve various impartial experts throughout the research process: ethics committees, monitors and potentially even governmental authorities. The guidelines (together with the research protocol) provide the impartial experts with the set of rules against which the researchers can be hold accountable. So-called 'sponsors' defined in these guidelines are the main responsible legal entity in a clinical trial to ensure final accountability of all the responsibilities outlined in the document. The document also provides impartial experts with the tools that help them to execute their roles, e.g. audit trails, monitoring reports or training logs.

As specially trained experts, these roles add to the expertise available in the research team. They are also impartial in the sense that the actions and judgments of impartial experts are often constrained by institutional principles and explicitly articulated values (e.g. principles of research ethics) and the institutions of which these experts are part of have mechanisms that can hold these experts accountable in the case of wrongdoing. Often, governance rules of such institutions require the absence of financial ties, or other conflict of interests. Being bound by their institutional roles, these experts increase the impartiality of the research undertaking. In line with John's two-premise model of trust in science, these guidelines define the epistemic standards for good data (e.g., 'Principal Investigators need to be adequately educated') while also enabling the enforcement of this standard (e.g., through impartial ethics committees that verify whether Principal investigators are adequately educated). The impartiality of randomisation and blinding procedures' plausibly adds to the acceptability of th eRCT as the main method to generate acceptable evidence by all parties, as argued by Teira. Randomisation and blinding alone, however, cannot ensure that data collection also implements these features with competence and integrity. For that, we need to turn towards the functioning of the community. Clinical practice guidelines indicate that a focus on researchers alone would be too narrow to capture how well the community is organised. Rather we need to extend the focus and include stakeholders, such as ethics committees and monitors, in this well-ordered community to understand how evidence produced by this community can be trustworthy[22].

[22] The sociological literature has repeatedly shown that these requirements that I am sketching here are not always fulfilled. Quite the opposite might be true: Employees of regulatory agencies often do have conflict of interests; they have close ties to people working in pharmaceutical companies and it is not uncommon that they circulate between jobs at the agency and regulators (the 'revolving door'). In a similar vein, Hauray argues that the scientific ideal of regulatory decision making – the ideal that decisions are made solely based on scientific evidence – is only a pretence. See Hauray (2017). This literature shows that trust even in the larger biomedical research community can be misplaced. What I am interested here, is the contrast between clinical trials and real-world data. If one holds that trust in biomedical research is always misplaced, this is a fortiori the case for real-word data.

1.2. Two arguments against the trustworthiness of real-world data

The previous section provided an understanding of the enablers of trustworthiness in clinical trials. In this section, I explore how these enablers could be employed in the production of real-world data to similarly ensure its trustworthiness. The question I explore is how the practice of data quality assessment could be embedded into the overall well-ordered clinical research community to make the practice of quality assessment trustworthy. To begin with, it might surprise that studies using routine data require none of the instruments or impartial experts to be involved that are so commonly employed in clinical trials. Recall the first study submitted to the FDA for the indication extension of Prograf discussed in the previous two chapters. Because the study was observational and built on retrospective data, a data use agreement between the SRTR and the researchers was sufficient. As indicated by the authors of the published study, 'The SRTR is made available under a Data Use Agreement to external researchers. No Institutional Review Board, Independent Ethics Committee, or Competent Authority approval was required for this analysis.' (Erdman et al. 2022, p. 1234). In most countries, studies using routine data do not require ethics approval and the research protocol is optional. Registries or EHR databases are generally not equipped with audit trail functions and do not require monitoring. A review of about 50 international routine health data–sharing initiatives in Europe found that none of the initiatives used software that was suitable for external audits (Bernal-Delgado et al. 2022). Moreover, the expertise required to handle repurposed data is remarkable, yet experience with the requisite skills for routine studies is still lacking. Leonelli mentions that consortia provide an important platform to discuss what counts as expertise in the field of bioinformatics (Leonelli 2016, chapter 2). For clinical trials, the skills and competencies requires are quite well understood and have been systematically documented in form of 'core competencies frameworks (Glaettli et al. 2022).

The community that performs these types of studies differs substantially from the community in which the production of evidence from clinical trials is embedded. Generally, none of the requirements to ensure the trustworthiness of clinical trial data are employed for studies based on routine data. In some cases, this general picture is more nuanced. Data that is submitted to regulators certainly undergoes critical examination by these regulators after it has been produced. Moreover, following its real-world evidence guideline, the FDA recommends pre-

planning the study with a protocol, and in most cases the FDA would probably review the protocol and provide critical feedback to minimise controversial value judgements (US Food and Drug Administration 2021b). Moreover, in some countries, ethics approval is required for studies with routine data. Yet even in these cases, the critical scrutiny of such data still falls behind the well-established standards employed for clinical trials.

This argument hinges on the contingency of the status quo. Yet this situation is busy changing. Stakeholders are well aware of the impactful consequences of the lack of instruments to ensure the trustworthiness of data. Many stakeholders are thus advocating for the deeper involvement of impartial experts and the use of related instruments in real-world studies. For example, reporting guidelines recommend the pre-registration of such studies (Kwakkenbos et al. 2021). Large initiatives advocate for the transparent documentation of data curation and verification practices of large databases (Bernal-Delgado et al. 2022). Data sharing initiatives advocate that data infrastructures should be monitored or that EHR systems should be equipped with audit trail functions (Monitoring Platform of the SCTO 2023). There are various ongoing activities to embed this research into the extended and trustworthy social community. These developments may imply that real-world evidence could eventually attain the same level of acceptance and trustworthiness as clinical trials due to the increased regulatory standards. However, two key obstacles remain: increasing the regulatory standards might not be desirable and it might not be feasible. I elaborate on the first obstacle now, with the second discussed later in this chapter.

Why should increased regulatory standards for real-world evidence not be desirable? The primary reason is straightforward: every additional requirement imposed on research that employs routine data diminishes the main epistemic value of such studies – its power. Real-world data enables researchers to use the limited resources more efficiently and answer more questions more rapidly, because real-world data is cheaper and quicker to obtain than experimental data. Wilholt coined the term epistemic 'power' to refer to this desiderata, that is, the ability of a method to produce results at a high rate with limited resources. He has further shown how power generally trades-off with reliability of evidence (Wilholt 2013; for a discussion see Chapter 5) I suggest that another trade-off emerges namely between the data's power and its trustworthiness. Undeniably, a major factor contributing to the high cost and prolonged duration of clinical trials is the slow, highly demanding and costly regulatory process in which they are embedded. Some scholars argue that the complexity of these regulatory requirements has led

pharmaceutical companies to depend on the expensive services of so-called contract research organisations and gave rise to a whole research infrastructure industry (Collins et al. 2020). In contrast, research with routine data is considered a 'last resort' that is easily accessible even to low-resourced research environments. It is clear that increasing the regulatory requirements would deprive routine data of this advantage.

I agree with Wilholt that power of evidence is a desirable value of real-world evidence and hence data. I argue in Chapter 5 that this desideratum could substantiate and partially legitimise the epistemic shift towards using real-world evidence. I further argued that the epistemic power of real-world evidence is already overestimated. Hence, it should be a high priority not to further diminish the data's power. Balancing the trustworthiness of data is a delicate task: It is crucial for ensuring the community's acceptance of evidence, yet it must be done without compromising the data's power, since that power is a primary source of legitimacy of the epistemic shift in the first place.

Trustworthiness and reliability of data are closely linked, yet they come apart. Trustworthiness indicates that we have reason to believe that the data has been acquired with the necessary expertise and has been reported truthfully and handled with uncontroversial value judgements. The first two requirements for the data's trustworthiness are prerequisites for the data to be reliable. However, conversely, data can be reliable without being trustworthy if we are not given any good reason to think so. Reliability and trustworthiness come fully apart regarding the issue of controversial value judgements; data can be reliably collected even if highly controversial criteria are employed. In other words, even if real-world data are sufficiently accurate to be 'fit-for-purpose' as required by data quality frameworks, they might not be trustworthy – in the sense that we cannot rule out that the data was fabricated, incompetently handled, only partially reported or collected based on controversial value judgements. Counteracting this threat comes at the price of the data's power, depriving us of the main reason for using such data to begin with. With the trade-off between the data's trustworthiness and its power we might be caught between a rock and a hard place.

2. Trustworthiness of data validation

So far, I have substantiated the view that the clinical trial community is well-ordered, and the evidence produced by this community is trustworthy – because the community includes impartial experts whose involvement is mandatory by law.

I provided a factual and normative argument against the idea that routine data research *is* or *should* be well-ordered in the same sense. In this section, I develop a third conceptual argument casting doubt on the idea that routine data *could* be well-ordered in the same sense. This third argument builds on methodological features of data validation, which play a central role in supporting claims about data's fitness-for-purpose according to the FDA guidelines for using real-world data for regulatory submissions (US Food and Drug Administration 2021b).

To illustrate what I am concerned with here, I use two claims made by Mehra and colleagues as an example. To support the fitness-for-purpose of the data in the Surgisphere publication they claimed that

> 'Collection of a 100% sample from each health-care entity is validated against financial records and external databases to minimise selection bias.'

and

> 'Verifiable source documentation for the elements include electronic inpatient and outpatient medical records.' (Mehra et al. 2020)

In Chapter 5, I explored methodological and epistemological issues of such claims. The questions I address in this chapter are as follows: 'Could the production of such claims be embedded in the larger well-ordered clinical research community to make them trustworthy? If so, how?'

I hold that there are two methodological features of data validation procedures that make it difficult or even impossible to embed the production of such claims into the larger well-ordered clinical research community. These features are deep local contextuality and high value-ladenness. These characteristics present considerable obstacles for identifying impartial experts and defining clear rules to which the community can be held accountable. With the argument in this section, I contend that the combined influence of local contextuality and the inherent value-ladenness of these practices might ultimately erode trust in data.

2.1. Local contextuality and value-ladenness of data validation

I noted in Chapter 5 that data validation practices require skilful and knowledgeable handling by experts. This was evident in the example of the assessment of the SRTR, where the researchers missed an essential quality issue in the data that rendered their primary analysis uninterpretable. It was mostly mere

luck that the problem was so clearly visible in the result and the researchers could act accordingly. Here, I deepen this line of argument by showing that data validation not only requires expert knowledge about the research context but deep local knowledge and various value judgements. If a scientific practice is *contextual,* this means that its epistemic goodness depends on the context in which it is performed or for which it is being used. Thus, doing the right epistemic thing requires tailoring scientific practices to these contexts. The corollary is that contextual factors may constrain the range of what counts as the right epistemic practice. In some sense of 'context' this is the case for many scientific practices. Clinical trials, for example, distinguish between many different medical contexts, e.g., types of diseases or different classes of treatments. Depending on these aspects, some outcomes might be considered informative while others are not. Tailoring scientific practices to the different contexts requires medical background knowledge; generally, people trained in the relevant medical specialty are competent experts to evaluate any given practice within its context. The contextuality of data validation runs deeper. Data quality depends not only on the type of medical context or the study question at hand but also on features that are local to a particular healthcare setting or database. Examples are local medical practices or local data handling practices that are not generally shared but need to be acquired by professionals by closely engaging with this local context. Leonelli has convincingly shown that data curation similarly relies on skilful handling of data and knowledge about the local contexts in which data have been generated. As part of their job, data curators report such local contexts in form of metadata (Leonelli 2016). The following elaborations are mostly show that the same is true not only for the curation of data but also of its validation.

To support this claim, I briefly summarise how data validation against an external reference standard works and point towards various sources of local contextuality. (My detailed discussion appears in Chapter 5 of this thesis.) Validation of data against an external reference standard involves three steps: classification of the data, construction of a reference standard dataset and quantification of data quality against the reference standard. The goal of a data validation study is to quantify how well an algorithm performs in capturing all – and only – the true events the researchers are interested in. In the simplest case, the algorithm simply translates the medical term into the semantic standards used in the database. For example, to classify the clinical event 'COPD', the algorithm extracts all instances of the ICD Code known as J44 and related codes. A subset of the resulting classification is then compared with a reference classification known

to be (more) accurate to estimate the accuracy of classification in terms of measures known from diagnostic tests, such as sensitivity or positive predictive value. An acceptable classification should at least account for reporting variability within the database; e.g., where different semantic standards are used. More complex algorithms can also account for various misrepresentation problems in the data, including reporting errors, errors of misdiagnosis or even measurement errors. For example, to classify a patients' record as an instance of COPD, an algorithm can be optimised to avoid false positive classifications of events as COPD that were none, by requiring the presence of commonly administered treatments for COPD together with the diagnostic code for COPD. Or an algorithm can be optimised to avoid false negative classifications of COPD by requiring broad symptoms rather than diagnosis. Validation studies usually compare different options and evaluate which one performs best. The second step is to create a reference standard against which the classified data can be validated. Common practices for constructing a reference standard include sending surveys to practicing physicians to ask them to confirm/disconfirm the classification, collecting original source data (such as laboratory reports) or using another routine data source (such as financial claims). I argue in Chapter 5 that the main methodological issues arise with this second step, because reliable and valid reference standards are hard to obtain. The third step is to compare the data that needs validating against this external reference standard and quantifying the difference in terms of measures well-known form diagnostic test such as specificity and positive predictive value.

Local expertise is required at several instances. To begin with, constructing a classification algorithm certainly requires familiarity with semantic standards and data structure that are local to a particular database. It also requires general medical knowledge as well as familiarity with local medical practices. One study constructed algorithms to classify the event 'acute exacerbations of COPD' and used medical knowledge about common symptoms of the condition as well as commonly prescribed medications to treat the condition. In cases where medical practices follow local standards that diverge from more widely shared medical knowledge, familiarity with the local practices is essential for constructing well-performing classification algorithms. Overall, success in algorithm construction is contingent upon a deep understanding of the local intricacies in both data curation and medical practices.

Choosing a suitable and feasible reference standard also requires local expertise about the standard's strengths and weaknesses. For example, in a widely used reference standard, researchers send out questionnaires to the practicing

physician of a subsample of patients, asking the doctor to confirm or disconfirm the data. The main problem of this procedure is its circularity, because physicians might 'confirm' the classification by looking up the very same data that is being verified. To judge the validity of such a reference standard, one needs to know whether physicians have an independent source of information available in this local setting. Using another routine data source as the reference standard requires local knowledge about the drivers of data quality in this reference standard. For example, a validation study used the HOB-DB as a reference standard to validate data on ethnicities in EHR data. The team relied on the high quality of the HOB database, based on their local knowledge about data-collection processes, funding incentives etc. Such knowledge was available to the team because one of its members was an employee at the hospital that curated the database (Kahn et al. 2010). Because of the local contextuality of the reference standard, data quality measures should be understood in relation to the specific factors that affect data quality in the reference standard. A positive predictive value of 95% might be good enough if the validation procedure lacks circularity but might be insufficient otherwise. Hence, these precise quality measures are barely informative unless one possesses local knowledge about potential pitfalls in the assessment.

Validating data against an external reference standard also involves decisive value judgements. One such judgement concerns the choice of a reference standard. Since generally all reference standards come with different epistemic risks, determining which reference standard is the best is not just a matter of making the right choice in a purely epistemic sense – but making the right choice in the sense of balancing the various risks aligned with non-epistemic value judgements. The most obvious value judgement involved in data validation is to determine a threshold for successful validation. This requires first determining which measures are most important in a certain case, e.g. specificity in combination with the positive predictive value. Second, one must set a threshold for these measures, e.g. a positive predictive value of 95%. Undoubtedly setting such a threshold depends on whether one is more concerned about false positives or false negatives. The medical context can put constraints on the range of acceptable value judgements, and the FDA's guidelines provide a few hints about this issue. For example, for rare or infrequent outcomes, the FDA recommends achieving high sensitivity and specificity. Yet, the question remains how high is high enough? The general recommendation by the FDA is the following: 'Some misclassification might be tolerable in some studies when the presence of misclassification is not expected to change the interpretation of results' (US Food and Drug Administration 2021b,

p. 11). In rare cases, the impact of misclassification on the results can be quantified. Yet, even in those cases, the quantification cannot eliminate the need for evaluative judgements about the quantitative threshold at which the interpretation of the results would change. Moreover, such quantification itself clearly includes value judgements. The idea is to recalculate the results of the clinical study under slightly varied assumptions to get an idea of how the different assumptions impact the result. This approach was used in the validation of the data that supported the FDA approval of Prograf for lung transplant recipients. The researchers recalculated the effect size of the treatment under the modified assumption that all patients whose outcome data was missing would have had an undesirable outcome. This assumption provided the researchers with the most conservative estimate of the treatment effect. The decision that this conservative estimate is the right assumption to make is a value judgement; the researchers accept the risk that the missing data is judged to be non-negligible, although the most conservative assumption might not be the most plausible. Hence, we cannot escape the need to make value judgements when setting quality thresholds, even if those judgements are based on quantitative information.

Validation of a single data item contributes only little towards an overall judgement about the data's fitness-for-purpose for a certain study. Assessing the overall fit-for-purpose of a dataset for a study requires to repeat this process for all essential data item, including diagnosis, treatment, safety and efficacy outcomes, eligibility criteria or confounding variables. How many of these variables must be validated before the data can be judged fit-for-purpose? According to the FDA, this point depends on how much uncertainty one is willing to live with:

> Overall, the required extent of validation should be determined by necessary level of certainty and the implication of potential misclassification on study inference. (US Food and Drug Administration 2021b, p. 21)

Moreover, 'data quality' is a multidimensional concept that covers the data's accuracy, completeness, relevance and more. All the study variables should be evaluated for each of these dimensions, and every dimension for every data variable will involve the problems of local contextuality and value-ladenness. Hence, there are not just a few but almost countless value judgements involved in the evaluation whether data is fit-for-purpose. These considerations reinforce concerns about the method's trustworthiness. Both, value judgements and local contextuality make the method highly susceptible to all sources of mistrust: lack of expertise, fraud or controversial value judgments. The solution that clinical trials employ is the

involvement of impartial experts who assure that everyone plays by the rules. The question I ask in the final section is whether this solution could be employed to ensure the trustworthiness of data validation.

2.2. A third argument against the trustworthiness of real-world data

The argument I develop in this last section is that these characteristics run so deep that they present veritable obstacles for embedding the production of claims about data quality into the larger well-ordered clinical research community to make them trustworthy. The claims I am interested in are claims of the sort: 'The SRTR database is fit-for-purpose to support the effectiveness of Prograf for lung transplant recipients', and 'Data for the primary outcome have been validated against financial records and deemed sufficiently accurate with a positive predictive value of 95%'. To recall, clinical trials that are conducted according to the rules of GCP can be seen as generating data that is trustworthy because they involve various impartial experts throughout the research process. Moreover, GCP guidelines and the research protocol provide these experts with a set of standards to which the community can be hold accountable; they also establish a standardised set of tools to execute roles such as audit trails, monitoring reports or training logs. This prompts two crucial questions: Who counts as an impartial expert? What are the rules everyone can be held accountable to in data quality assessment practices?

I begin with trying to identify impartial experts. Tailoring clinical research practices to the different contexts generally requires medical background knowledge. In this case, people trained in the relevant medical specialty are qualified experts to evaluate the practice at hand within its context. What expertise is required to tailor data quality assessment to the local contexts of its production? The most obvious answer is to involve someone who is knowledgeable about the local context. Leonelli has shown, that most data infrastructures have professional data stewards or curators whose job it is to clean, annotate or transform the data and make it available to others for secondary use (Leonelli 2016, chapter 2). After some years of experience, professionals in these roles would count as experts regarding the data in the database. Hence a potential solution could be to make it mandatory to involve such local experts into the quality assessment of routine data. Leonelli discusses this option and notes that this form of assessing data quality by data curators would mean to put a lot of responsibility to data curators and that 'this type of quality assessment is only as reliable as the curators in charge' (Leonelli 2017b, p. 4). She sees this option particularly problematic in cases where data

curators are disconnected from the community of data users or where datasets become so large that they cannot be manually handled by data curators. I agree that the expertise of data curators might pose a problem particularly because such expertise is yet not well-defined. Moreover, assessing the data's fitness-for-purpose requires both, knowledge about the data and about the research. Data curators know their data but they are rarely medical experts.

The greater concern, however, arises about whether these professionals also qualify as impartial. Data curators are at least partially responsible for the quality of their data and therefore might have an interest in overlooking potential quality issues. Moreover, data infrastructures might financially depend on collaborations with research projects that use their data. In that case, employees would experience a classical conflict of interest when assessing whether their own data is fit-for-purpose to be used in such a collaboration. An alternative approach could involve obtaining approval from an impartial ethics committee for a planned data validation method and quality thresholds. Although the ethics committee would qualify as impartial, they would not qualify as experts, because they do not possess the local knowledge required to evaluate the appropriateness of the measures. Hence, there seems to be a general tension between both desiderata – being an expert versus being impartial – to qualify as an impartial expert who could execute or verify data quality assessments. Without the involvement of impartial experts, however, the community is not well-ordered. Tempini and Leonelli present an interesting case study of information security practices at Secure Anonymised Information Linkage (SAIL). Within this infrastructure, the staff was divided into two roles they call 'infrastructure-facing' and 'research-facing'. Only the research facing analysts are involved in particular research projects which allows the infrastructure-facing experts to remain as free as possible from potential conflict of interests (Tempini and Leonelli 2018). For data infrastructures that are large enough and well-founded perhaps such divisions of roles could provide a partial solution to the problem of trustworthiness.

The second obstacle to embed data quality assessments into the well-ordered community is that we lack suitable rules to which everyone could be hold accountable. The variability in data quality practices is considerable and it has been repeatedly noted that this makes it difficult to define international standards (Leonelli 2017b) and that this hampers progress in the field (Illari 2014). Moreover, because the goodness of quality assessment practices depends so strongly on local contexts, it seems impossible to formulate rules that apply to all contexts. Such rules must be so general that they are almost idle. I consider here

the convention 'data must be validated against an external reference standard'. Because this convention allows for so much variability and necessitates countless value judgements, following this rule is insufficiently informative. The most specific propositions that guidelines might define is set of bold heuristics such as 'Claims data on prescription tend to be more accurate than EHR data and could be used as a reference standard to validate treatment data.' Such recommendations can provide some guidance, but they cannot be binding, because they might not hold in a particular context. Hence, it seems difficult if not impossible to establish a set of epistemic conventions that coordinate judgements about the data's fitness-for-purpose. Yet without such rules, neither of the two premises for trust can be fulfilled: Neither can such conventions be broadly acceptable to the community, nor can we claim that adherence to these conventions is the best explanation of a claim if we simply to not know what these conventions entail. Hence it is no longer the case that data that has been deemed fit-for-purpose according to the standards of one group of researchers must be accepted as fit-for-purpose according to the standards of another group of researchers.

The alternative is to fully entrust local experts to judge the relevance of general heuristics for their particular context on a case-by-case basis. Leonelli clearly holds that the contextuality and diversity of data quality practices calls for embracing a localised approach to data quality and a reflexive exercise about its underlying assumptions. (Leonelli 2017b). Although adopting a contextualised approach to data quality assessment is a viable option, fully embracing it carries profound implications. It means giving up on shared epistemic rules about what counts as good data and instead entrusting the definition of such standards on a case-by-case basis to the few local experts who are competent for this job. Such a locally contextualised approach to data quality also means giving up on the accumulation of shared experience in working with a particular set of rules – rules on which the current good functioning of the global research community is based.

It appears that RCTs lend themselves particularly well to the current rule-based operationalisation. Randomisation already minimises the number of judgements needed. The practice of verifying research data against source data is an unambiguous rule that anyone, even without medical expertise, could follow. The control over data collection makes it possible to formulate conventions about the precise data that is required for each type of investigation. For example, guidelines can define the type of outcomes that are acceptable as proof of effectiveness for every type of disease. This is not to say that RCT do not require medical expertise or background knowledge to be designed, they certainly do.

They also involve value judgements. For example, researchers have to judge the acceptable drop-out rate in an experiment. Such judgements require medical expertise, yet they are by far less dependent on local contextualities.

The turn towards a contextualised approach of data quality implies a radical shift. Currently the community coordinates its judgments on a global scale, based on a set of highly legitimate and legally binding rules. Adherence to these rules is overseen by a network of impartial experts throughout the research process. Embracing the locality of real-world data means to settle for a set of unbinding heuristics whose relevance can only be evaluated by a few highly specialised local experts with uncertain impartiality. Consequently, the epistemic shift that is busy happening is not only a shift from one set of rules to another one. Rather, we might be witnessing a more radical shift from the paradigm of regulatory oversight to a paradigm of trust in local experts. The aim of turning toward a fitness-for-purpose approach is precisely to abandon the 'one-size-fits-all' approach and embrace the role of expert judgments. The future will show whether the clinical research community and those reliant on the knowledge it produces are prepared to embrace such a radical shift.

Conclusion

The uptake of real-world evidence into the norms and practices of clinical research will transform knowledge about medical interventions in various ways. I explored how pragmatic trials create conceptual shifts in our medical knowledge including a shift towards broader medical interventions (Chapter 2). I argued that such conceptual shifts will also be fairly common for other concepts because of the many limitations that are prevalent in real-world data (Chapter 4). I further showed that some of these limitations will bring advantages like an increased focus on outcome measures, comparator treatments or populations that are clinically and practically relevant (Chapter 3), while their limitations might introduce controversial compromises like some of the compromises made in the assessment of Prograf (Chapter 4). Moreover, I argued that real-world studies can provide a valuable perspective because they yield effect estimates that can be seen as more 'realistic' (Chapter 3).

My study highlights that issues about data quality are becoming another paramount topic in medical research. Real-world data shifts the current understanding of data quality from adherence to the rules of GCP towards the idea of 'fitness-for-purpose' (Chapters 4). Data quality assessments are very heterogeneous practices, and I argued that their reliable assessment is difficult to achieve (Chapter 5). The deep local contextuality and value-ladenness of these practices might even bring about new foundations of trust in medical knowledge because they are no longer easily verifiable by impartial experts (Chapter 6). To prepare for the changes ahead, it seems paramount to start exploring new mechanisms of accountability, to further advance a theory of medical interventions and to reflect deeper on the interdependence between purpose and epistemic notions such as validity, bias, or data quality.

With the advent of real-world evidence into the regulatory realm, evidential standards transition from a hierarchical and rule-based evidence paradigm towards a more contextualised approach where evidence is flexibly defined depending on the context of use. My study of pragmatic clinical trials has demonstrated how even a subtle shift of the purpose of a clinical trial can change what counts as bias (Chapter 2). The discussion on the notion of fitness-for-purpose has shown how flexible the notion is to incorporate various commitments about the necessary conditions to achieve a purpose (Chapter 4).

The advantage of a contextualised approach is that it can recognise the multidimensionality of 'quality' of evidence, can highlight those dimensions that are most essential in a particular context, account for available resources and resource constraints and carefully tailor the choice of methods to these circumstances. Both case studies illustrate these strengths. The Salford Lung Study has shown that pragmatic trials that are conducted under routine care conditions can yield new insights about the effectiveness of treatments that are relevant for making treatment decisions (Chapter 2). The approval of Prograf as an immunosuppressant for lung transplants demonstrated its effectiveness with remarkable efficiency and low risk of making an erroneous decision (Chapter 4). However, putting the approach to use comes with serious risks. Most serious is the risk of unreliable data quality assessments. Since the notion of 'fitness-for-purpose' already makes an allowance for reduced data quality, tolerating unreliable data quality assessments is a risk that we should not be willing to take (Chapter 4). The deep local contextuality and value-ladenness of such practices make data quality assessments even more worrisome because this undermines the possibility of regulatory oversight by impartial experts. The new evidence paradigm might require new foundations of trust (Chapter 6). Where researchers will rely on observational designs, the problems of data quality will be accompanied by the old problem of confounding. I have argued that this problem can and should be avoided by preferring pragmatic trials over observational designs (Chapter 1). Nevertheless, the high flexibility of the contextualised approach introduces considerable leeway for the research community that easily lends itself to misuse.

Real-world evidence comes with great promises. I have studied two of them: the promise that such evidence is informative of clinical practice settings and the promise that it uses limited resources more efficiently. I agree with the advocates of this development that these are real concerns that could, in principle, provide a substantial epistemic justification for a change in evidence standards. However, my investigations have shown that we should take a more critical stance towards these promises. 'Real-world evidence' is first and foremost a clever rhetorical trick distracting its users from the many quality issues that such data entails. I have argued that properly conducted real-world studies, like pragmatic trials, can provide what I called 'realistic' effect estimates because they are conducted under natural and non-ideal conditions. However, this rational falls short of the widely hold intuition that such evidence is 'widely generalisable' or informative about patient level treatment decisions. Rather it might provide a glimpse into the effects of medical interventions in interaction with 'imperfect users' (Chapter 3). 'Secondary

use data' appears to be the epitome of a powerful research approach and constitutes the second promise of this development. Yet, I argued that the production of real-world data also distributes the costs of producing medical knowledge towards public third parties, including healthcare providers, regulators, governments and patients. Taking all the costs into account that accompany this development, it is likely that we overestimate the power of real-world data.

The diversity of real-world evidence comes with a greater risk still. Our understanding of the RCT gold standard is built on decades of experience and many lessons learned from failures in the past. In light of the sheer endless diversity of methods and data that real-world evidence brings along, it is questionable whether the community will ever reach a similar level of methodological understanding about its use. Increasing the diversity of evidence comes with the chance to look at treatments from different perspectives, but it fundamentally also comes with the risk that disconvergent evidence will multiply disagreement. The difference between pragmatic and explanatory trials is just one example of the increased diversity of evidence that we are going to see. Already in this case, understanding their complementary strenghts and weaknesses is a complex task. It has become common sense among opinion leaders in the field that real-world evidence and RCTs are not in competition but merely complement each other. The risk that such an attitude entails is that we multiply the evidence without a proper understanding *how* the different approaches complement each other.

I have tried to do pay justice to the multifaceted characteristics of the evidential evolution that awaits us. Nonetheless, many interesting philosophical questions and important scientific practices could not be addressed that present avenues for future research. Most importantly, this concerns a detailed study of the ongoing empirical efforts to replicate results from RCTs with real-world evidence. It would be interesting to study not only the success of these replication studies, but also their replication goals and the methods used. These empirical investigations hopefully yield insights into the sources of heterogeneity of evidence or the possibility to reproduce data quality assessments that could complement my philosophical investigation. Moreover, despite the breadth of my study on real-world evidence, the scope I covered is only a fraction of all the uses of this evidence. I focused on the use of such evidence for the study of effectiveness of medical interventions, while I did not look at its use for the study of safety or other purposes, including hypothesis generation, cost-effectiveness studies or the monitoring of quality in routine care. Finally, at the end of writing of this thesis, the ICH has

published the third revision of the GCP guidelines, containing new principles for data management for research with secondary use data. Hopefully, a closer look at these guidelines might reveal a solution to what I exposed as the problem of trustworthiness of real-world data.

Annex: A deflated definition of the internal–external validity distinction

Many arguments about the epistemic strength and weaknesses of RCTs in clinical research employ the notions of internal validity and external validity. Particularly the epistemic superiority of RCTs has been criticised, on the ground that such designs lack external validity. That is, it is generally not possible to generalise evidence from RCTs to other settings and populations (Deaton and Cartwright 2018). Similarly, pragmatic trials are thought to increase external validity at the expense of internal validity; this point is often mentioned when discussing the epistemic merits, problems and tensions of these designs (Godwin et al. 2003). In this Annex, I use lessons learn from the strengths and weaknesses of pragmatic-explanatory distinction to substantiate the internal-external validity distinction and defend the definition I used in Chapter 2.

Conceptual and theoretical reflections on the internal–external validity distinction are dominated by vague claims that conflate several aspects of these concepts. In the scientific literature, external validity and internal validity are usually introduced simultaneously, and the former is defined in close relation to the latter. For example, Patino and Ferreira (2018) describe internal validity as 'truth in the study' and external validity as 'truth in real life'. In the philosophical literature, Guala clarifies that internal validity is 'a problem of identifying causal relations' and 'is achieved when the structure and behaviour of a laboratory system ... have been properly understood by the experimenter'. By contrast, 'external validity involves an inference to the robustness of a causal relation outside the narrow circumstances in which it was observed' (Guala 2003). Schram states that internal validity 'will yield results that are robust and replicable' (Schram 2005).

Most importantly, the weak conceptual scaffold has fostered conflicting views on the relation between internal and external validity and how to prioritise them. A popular view is that internal validity *trades-off* with external validity; i.e., increasing the external validity comes at the expense of decreasing the internal validity. Cartwright called the trade-off between internal and external validity a 'well-known truism' (2007). In her view, the deductive rigour of any experiment is gained through rather demanding assumptions, which in turn necessarily limit the scope of the conclusion. Others have attempted to explain the trade-off as a tension

between 'artificiality' and 'reality' (Schram 2005). Another popular view is that internal validity occurs epistemically prior to external validity. For example, Guala influentially argued that it does not make sense to bring up the question of external validity unless one is first confident about internal validity, because internal validity is epistemically antecedent (Guala 2003).

It is mostly in the philosophical literature on experimental economy that we find attempts to clarify these two mainstream perspectives and whether they are compatible (Jimenez-Buedo and Miller 2010; Schram 2005; H. Chytilová and R. Maialeh 2015; Persson and Wallin 2015). However, these contributions mostly work with implicit definitions of the terms that make decisive presuppositions. I want to make explicit some of these presuppositions and propose an account of internal and external validity that relies on a threefold characterisation of the concepts. In this view, a trade-off between the two can occur contingently, because some experimental properties are incompatible practically. The view that internal validity is prior to external validity, on the other hand, is based on an equivocation of the terms.

I defend this deflated explication of the internal–external validity distinction as preserving its usefulness. The usefulness of the distinction lies in its role for assessing, criticising and designing experimental designs. It is in this role that the distinction fostered various quality assessment tools and a collection of empirical evidence about factors that impact the validity of experimental designs. Such work is in line with the distinction's historical origin when Campbell first introduced it in 1957. Thus, equally important to the meaning of these terms is to clarify what is at stake if we attribute their presence or absence to an experiment. Hence, I am aware that these concepts are used in several ways that depart from my proposal. I believe the meaning I am defending here preserves the distinction's usefulness.

Substantiating the internal–external validity distinction

The general concept of validity in experimental research has been systematised by attempts to distinguish between different types of validity. Most prominently, in the late 1950s Campbell introduced a typology of validity: First, he distinguished between internal and external validity, and later he extended the typology by adding the terms 'statistical conclusion validity' and 'construct validity'. I am in no position to propose a comprehensive exegetical analysis of Campbell's work and the evolution of the concepts he introduced. Here, I make a few comments about how my proposal relates to and diverges from its origins in Campbell 1957.

I propose an explication of internal and external validity that relies on a threefold characterisation of the concepts: each type is characterised by a) a type of proposition that can be inferred from the experimental results, which is b) supported in virtue of a cluster of properties of an experimental design that c) perform a corresponding epistemic role.

For internal validity, the following explication holds:

An experimental design is internally valid if an inference from its results to the proposition that the 'result was caused by the intervention' is supported in virtue of its experimental properties.

This is the corresponding explication for external validity:

An experimental design is externally valid if an inference from its results to the proposition that 'the experimental results will generalise to this target' is supported by virtue of its experimental properties.

These explications are deflated in three ways. First, the content of the propositions supported by either quality attribute is modest. E.g. internal validity inherits the notion of the intervention from the experiments, and external validity only requires that the results generalise to a specific target setting. Second, validity is an attribute of experiments that pertains to them in virtue of their experimental properties. Thus, attributing the absence of either internal or external validity to an experiment does not imply anything about the overall evidential support of the proposition outside the experiment. In other words, neither type of validity is necessary to gain knowledge. Third, neither internal nor external validity require establishing the truth of the propositions but only lends some support to the proposition. This amounts to saying that experiments with such quality attributes are not sufficient to establish the truth of the propositions; sometimes they might not be sufficient to establish any strong support. Hence, I argue that the usefulness of the internal–external validity distinction lies in its use as a tool to assess and characterise experiments. A deflated notion highlights the usefulness of the distinction because it acknowledges its limits, as both quality attributes are neither necessary nor sufficient for knowledge generation. I now elaborate on all three aspects.

Propositional content

Internal validity supports causal claims, so much is common sense. When thinking of an internally valid experiment, many people have in mind an experiment that licenses a causal inference to a single causal variable. Such usage became so commonplace that in his later work Campbell intended to relabel his internal–external validity distinction to free it from such connotation: '... the term internal

validity now means similarity to the pure treatment (rule-of-one-variable), fully controlled, laboratory experiments. Since that is not what we had in mind, we need to try again, with new terms' (Campbell 1986 - Relabelling internal and external validity). In such an understanding of internal validity, the effects of multicomponent interventions, therapeutic actions or even holistic interventions – which are common in CAM – cannot be supported by internally valid experiments, by definition. However, in that, case internal validity is no longer worth pursuing for every experiment, as the accusation that an experiment fails to investigate a single variable is largely uninteresting. What is at stake when experiments are criticised for lacking internal validity is nothing less than the reliability of the causal inference, not the number of variables involved. Those who adopt such a notion of internal validity are willing to forego its evaluative import and the very reason why internal validity is usually considered indispensable.

In contrast, I hold that we should understand internal validity as supporting causal claims that relate the experimental results to the intervention. Thus, the validity attribute simply inherits the notion of the intervention from the experiment, which creates a conceptual interdependence between the validity attribute and the definition of the intervention. I elaborated this aspect when discussing pragmatic clinical trials in Chapter 2. If one wants to measure the isolated effect of the characteristic features of a therapy, then all other causal factors (e.g., the awareness of patients) have to be considered a distortion or bias in the treatment effect. If, however, the intervention of interest is a therapeutic action, a multicomponent therapy or even a holistic therapy, such factors are part of the intervention for which a causal relation has to be established.

External validity, on the other hand, is concerned with generalisation to a target setting. Here, the precise content of the subject and predicate are of great relevance. If one aims to generalise some results to all target settings, one requires from external validity what no single experiment can deliver. Another aspect is where the generalisation departs from. In my proposal, we depart from 'the experimental results'. As this notion is ambiguous, we need to clarify. It can mean at least three things: The measured difference by comparing groups or states with each other; a statement about the statistical significance of the measured differences; or even the causal interpretation of this difference. If one assumes that the generalisation departs from the causal interpretation of the results, one buys into the logical or epistemic primacy of internal validity by definition. If all that is required for generalisation is the statistical significance of the measured difference (whether causal or correlational), the logical or epistemic primacy disappears.

Hence, such a logical dependency seems uninteresting, as the preference for a certain order of inferential steps has been built into the definition of the terms. If what one is interested in are causal claims that generalise to a certain target setting, both to be established: That the causal conclusion holds and that the results are generalisable. There is no special order required in which these points are established. Hence, I assume that external validity has not built the causal interpretation into its definition, as this further reflects how these quality attributes are instantiated separately on the experimental design level.

An attribute of experiments

External validity and internal validity have been attributed to various things, including experiments, experimental results, inferences, inferred claims and even interventions. What is thought to be the bearer of validity makes a difference to how these terms are used. For example, among philosophers, it is not unusual to use the terms as in these examples: 'Knowledge of mechanisms can help with determining the external validity of an intervention' and 'internal validity is not necessary to gain causal knowledge'. However, these concerns are distinct from concerns of researchers who are trying to 'resolve the struggle between internal and external validity'. Researchers would use the terms in ways like this: 'We describe the design of the trials and the steps taken to deal with the competing demands of external and internal validity' and 'External validity is maximised by having few exclusion criteria and by allowing flexibility in the interpretation of the intervention and in management decisions.' Because the internal–external validity distinction is a tool for assessing and characterising the quality of experiments, these terms are attributes of the experiments themselves. Thus, an experiment is internally or externally valid by virtue of its experimental properties.

In the case of internal validity, it is widely acknowledged that a reference to the experimental design is necessary in order to say something about the internal validity of its results. Moreover, the experimental properties that relate to internal validity are well-known and widely studied: randomisation, blinding, adherence control, drop-out control and others. Yet, experimental techniques to ensure internal validity are many; they not only depend on the method type and the research goal but can be domain- or even subject-specific. The epistemic role that the control fulfils is well-founded and usually explained as the role of eliminating alternative hypotheses.

For external validity, the claim is less intuitive. The literature on extrapolation in the philosophy of science mostly uses the term 'external validity'

to denote the search for any kind of evidential support to generalise causal claims from all kinds of experiments. Here, 'mechanistic knowledge can support the external validity of a causal claim' might be a perfectly fine way to talk; extrapolating a causal claim means the same as establishing its external validity. However, in such usage, external validity is no longer attributed to the experiment but to the causal claim. It becomes an indicator for generalised knowledge in general. In my view, it is of little surprise that philosophers have claimed that the external validity terminology can do little work for them (e.g. Cartwright). It provides no means to disentangle the many different questions that are involved in extrapolation. Instead, notions such as 'causal robustness' and frameworks for extrapolation inferences do a much better job at establishing the scope of causal claims.

In my view, the quality attributes of an experiment do not imply or require saying anything about other evidential support, outside the experiment. Hence, the two questions of how to design an experiment to achieve maximal support for the generalisability of its results to a target, and what evidence can support that a causal claim holds in a different context, can be kept apart. If we think of validity attributes as pertaining to the experiment and not the proposition independent of any specific experiment, it is clear why internal or external validity are not necessary for supporting the proposition with evidence from *outside* the experiments. To illustrate, philosophers repeatedly argue that causal knowledge about medical interventions does not necessarily come from internally valid RCTs. They are right in claiming this – yet their point does not diminish the value of internal validity but only exposes its limits. Likewise, an externally valid experiment is not necessary to establish general causal claims, as we have various other methods to establish the scope of causal claims. Many of those methods could be more successful than considerations about the external validity of an experiment. However, if we follow the suggestions of some philosophers to drop the notion of external validity entirely, we will deprive ourselves of an important method to evaluate and criticise experimental designs. Notably, we could not articulate what is wrong with many randomised trials when they provide results that only hold within an artificially idealised experiment. While the properties that support external validity are generally less well-systematised than those supporting internal validity, pragmatic clinical trials have greatly contributed to develop such understanding. Hence, we should not nullify the difference between experiments with and without the following properties: random sampling or inclusive sampling, experiments that are embedded into ordinary processes or experiments that suffer from artificial effects that only occur in the experimental context (e.g. Hawthorne effect). There are

other examples beyond that list. For external validity, the experimental design supports local extrapolation of the results in a target setting, perhaps because it has optimised the experimental conditions to be similar to the target setting.

A final point is noteworthy. As an attribute of the experiment, validity can only play a justificatory role and comes without a truth condition. Thus, validity is not – as is sometimes erroneously claimed – a warrant or a promoter for truth or the discovery of facts. Take for example this widely quoted description of internal validity by Guala:

> [T]he result of an experiment E is internally valid if the experimenter attributes the production of an effect B to a factor (or set of factors) A, and A really is the (or a) cause of B in E. (Guala 2003, p. 1198)

Guala seems to say that internal validity is implied by a true belief, i.e., a given correspondence between the experimenter's belief about a causal proposition and a causal fact in the world. This idea demands a lot more from internal validity than the best experimental setup could deliver. Acknowledging the supportive role of validity instead yields useful consequences: Experimental results can be statistically, internally and externally valid – and still be false. In other words, validity is not sufficient to establish the truth of the propositions involved. Indeed, we can imagine plenty of things that can go wrong beyond the lack of internal validity. Such a justificatory role is also more in line with the notion that validity comes in degrees. Rather than being concerned with the truth simpliciter, internal validity is concerned with degrees of justification or support.

I believe that the reflections above support that the quality dimensions of an experiment can be assessed and judged independently of each other. That is, an experimental design can be evaluated for how it supports the applicability of its results to a target context, without having raised the question of whether it supports a causal interpretation of its results. Any other usage in this regard seems to undermine the very reason for drawing the distinction in the first place.

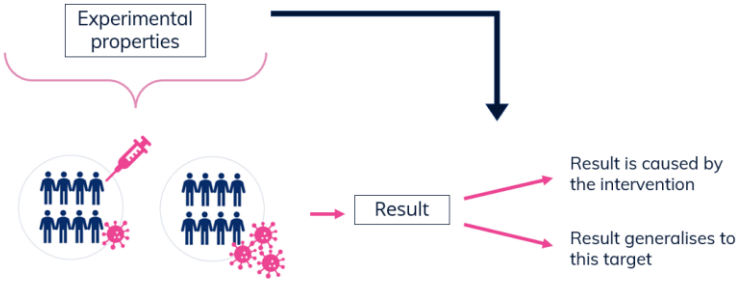

Figure 3: Experimental properties justify simultaneously two inferences from the observed results.

A contingent trade-off

The deflated understanding supports the view that there is a trade-off between the two although a contingent one. Cartwright called the trade-off between internal and external validity a 'well-known truism' (2007). In her view, the deductive rigour of any experiment is gained through rather demanding assumptions, which in turn necessarily limit the scope of the conclusion. Others have attempted to ground the trade-off in a tension between 'artificiality' and 'reality' (Schram 2005). While others have argued that there is no general tension between 'artificiality' and 'reality' have rejected the idea of a trade-off on these grounds (Jimenez-Buedo and Miller 2010; Schram 2005). From the reflections above, it follows that the trade-off simply emerges because some experimental properties that are useful to increase either type of validity are mutually exclusive with properties that increase the other type: blinding is mutually exclusive with the natural awareness of patients, flexibility in adherence excluded its control and the use of randomisation excludes the use observational data. Indeed, Campbell proposed the same idea:

> Both criteria are obviously important although it turns out that they are to some extent incompatible, in that the controls required for internal validity often tend to jeopardize representativeness. (Campbell 1957)

It is important to recognise that this is not a necessary trade-off in every instance, as there are means to increase the external validity that do not diminish the internal validity, and vice versa. The choice of a representative patient population can increase external validity without impacting internal validity. Similarly, randomisation can greatly increase the internal validity while only minimally impacting external validity.

References

Agustí, A., Teresa, L. de, Backer, W. de, Zvarich, M. T., Locantore, N., Barnes, N., et al. (2014). A comparison of the efficacy and safety of once-daily fluticasone furoate/vilanterol with twice-daily fluticasone propionate/salmeterol in moderate to very severe COPD. *European Respiratory Journal, 43*, 763–772. doi:10.1183/09031936.00054213.

Alfirevic, Z. (2023). *Assessment of trustworthiness has a significant impact on conclusions of Cochrane reviews,* Cochrane Colloquium 2023. Cochrane Collaboration, September, 4.

Altman, D. G. (1994). The scandal of poor medical research. *BMJ : British Medical Journal, 308*, 283–284. doi:10.1136/bmj.308.6924.283.

Andersen, J. R., Byrjalsen, I., Bihlet, A., Kalakou, F., Hoeck, H. C., Hansen, G., et al. (2015). Impact of source data verification on data quality in clinical trials: an empirical post hoc analysis of three phase 3 randomized clinical trials. *British Journal of Clinical Pharmacology (BJCP), 79*, 660–668. doi:10.1111/bcp.12531.

Andreoletti, M., & Teira, D. (2019). Rules versus standards: What are the costs of epistemic norms in drug regulation? *Science, Technology, & Human Values, 44*, 1093–1115. doi:10.1177/0162243919828070.

Arlett, P., Kjaer, J., Broich, K., & Cooke, E. (2021). Real-world evidence in EU medicines regulation: Enabling use and establishing value. *Clinical Pharmacology & Therapeutics*. doi:10.1002/cpt.2479.

Ashcroft, R. (2002). What is clinical effectiveness? *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences, 33*, 219–233. doi:10.1016/S0039-3681(02)00020-1.

Baetu, T. M. (2020). Causal inference in biomedical research. *Biology & Philosophy, 35*, 1–19. doi:10.1007/s10539-020-09760-4.

Balas, E. A., & Boren, S. A. (2000). Managing clinical knowledge for health care improvement. In Bemmel J & McCray A. T. (Eds.), *Yearbook of Medical Informatics 2000: Patient-Centered Systems. Stuttgart* (pp. 65–70). Stuttgart, Germany: Schattauer Verlagsgesellschaft mbH.

Bartlett, V. L., Dhruva, S. S., Shah, N. D., Ryan, P., & Ross, J. S. (2019). Feasibility of using real-world data to R replicate clinical trial evidence. *JAMA network open, 2*, e1912869. doi:10.1001/jamanetworkopen.2019.12869.

Bernal-Delgado, E., Craig, S., Engsig-Karup, T., Estupinan-Romero, F., Sahlertz Kristiansen, N., & Bredmose Simonsen, J. (2022). *European health data space data quality framework*. https://tehdas.eu/app/uploads/2022/05/tehdas-european-health-data-space-data-quality-framework-2022-05-18.pdf. Accessed 07.02.24.

Bluhm, R. (2017). Inductive risk and the role of values in clinical trials. In K. C. Elliott & T. Richards (Eds.), *Exploring inductive risk: Case studies of values in science* (pp. 193–214, Vol. 1). Oxford University Press.

Bokulich, A., & Parker, W. (2021). Data models, representation and adequacy-for-purpose. *European Journal for Philosophy of Science, 11*, 2–26. doi:10.1007/s13194-020-00345-2.

Borgerson, K. (2005). Evidence-based alternative medicine? *Perspectives in Biology and Medicine, 48*, 502–515. doi:10.1353/pbm.2005.0084.

Borgerson, K. (2009). Valuing evidence: bias and the evidence hierarchy of evidence-based medicine. *Perspectives in Biology and Medicine, 52*, 218–233. doi:10.1353/pbm.0.0086.

Borgerson, K. (2013). Are explanatory trials ethical? Shifting the burden of justification in clinical trial design. *Theoretical medicine and bioethics, 34*, 293–308. doi:10.1007/s11017-013-9262-4.

Borgerson, K. (2016). An argument for fewer clinical trials. *The Hastings Center report, 46*, 25–35. doi:10.1002/hast.637.

Bower, J. K., Patel, S., Rudy, J. E., & Felix, A. S. (2017). Addressing bias in electronic health record-based surveillance of cardiovascular disease risk: Finding the signal through the noise. *Current epidemiology reports, 4*, 1–12. doi:10.1007/s40471-017-0130-z.

Briel, M., Olu, K. K., Elm, E. von, Kasenda, B., Alturki, R., Agarwal, A., et al. (2016). A systematic review of discontinued trials suggested that most reasons for recruitment failure were preventable. *Journal of clinical epidemiology, 80*, 8–15. doi:10.1016/j.jclinepi.2016.07.016.

Broadbent, A. (2013). *Philosophy of epidemiology* (New directions in the philosophy of science). Basingstoke: Palgrave Macmillan.

Callahan, T. J., Bauck, A. E., Bertoch, D., Brown, J., Khare, R., Ryan, P. B., et al. (2017). A Comparison of Data quality assessment checks in six data sharing networks. *eGEMs, 5*, 8. doi:10.5334/egems.223.

Cambrosio, A., Keating, P., Schlich, T., & Weisz, G. (2006). Regulatory objectivity and the generation and management of evidence in medicine. *Social Science & Medicine, 63*, 189–199. doi:10.1016/j.socscimed.2005.12.007.

Cambrosio, A., Keating, P., Schlich, T., & Weisz, G. (2009). Biomedical conventions and regulatory objectivity. *Social Studies of Science, 39*, 651–664. doi:10.1177/0306312709334640.

Campbell, D. T. (1957). Factors relevant to the validity of experiments in social settings. *Psychological bulletin, 54*, 297–312. doi:10.1037/h0040950.

Campbell, D. T. (1986). Relabeling internal and external validity for applied social scientists. *New Directions for Program Evaluation, 1986*, 67–77. doi:10.1002/ev.1434.

Canali, S. (2020). Towards a contextual approach to data quality. *Data, 5*, 90. doi:10.3390/data5040090.

Carlisle, J. B. (2021). False individual patient data and zombie randomised controlled trials submitted to Anaesthesia. *Anaesthesia, 76*, 472–479. doi:10.1111/anae.15263.

Cartwright, N. (2007). Are RCTs the gold standard? *BioSocieties, 2*, 11–20. doi:10.1017/S1745855207005029.

Cartwright, N. (2009). What are randomised controlled trials good for? *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition, 147*, 59–70. doi:10.1007/s11098-009-9450-2.

Cartwright, N. (2012). Presidential address: will this policy work for you? Predicting effectiveness better: How philosophy helps. *Philosophy of Science, 79*, 973–989. doi:10.1086/668041.

Cartwright, N. (2017). *What's the use of pragmatic trials?* Department of Philosophy King's College London, January, 26.

CDER (FDA). (2021). *Multi-Discipline Review: Application Number: 50708s053, 50709s045, 210115s005*. https://www.accessdata.fda.gov/drugsatfda_docs/nda/2022/050708Orig1s053;%20050709Orig1s045;%20210115Orig1s005.pdf. Accessed 21 February 2023.

Chalmers, I., & Glasziou, P. (2009). Avoidable waste in the production and reporting of research evidence. *The Lancet, 374*, 86–89. doi:10.1016/S0140-6736(09)60329-9.

Clarke, B., Gillies, D., Illari, P., Russo, F., & Williamson, J. (2013). The evidence that evidence-based medicine omits. *Preventive Medicine, 57*, 745–747. doi:10.1016/j.ypmed.2012.10.020.

Clarke, B., Gillies, D., Illari, P., Russo, F., & Williamson, J. (2014). Mechanisms and the evidence hierarchy. *Topoi, 33*, 339–360. doi:10.1007/s11245-013-9220-9.

Claxton, A. J., Cramer, J., & Pierce, C. (2001). A systematic review of the associations between dose regimens and medication compliance. *Clinical Therapeutics, 23*, 1296–1310. doi:10.1016/S0149-2918(01)80109-0.

Collins, R., Bowman, L., Landray, M., & Peto, R. (2020). The magic of randomization versus the myth of real-world evidence. *The New England journal of medicine, 382*, 674–678. doi:10.1056/NEJMsb1901642.
Dahly, D. (2019). Out of balance. https://statsepi.substack.com/p/out-of-balance. Accessed 29 January 2021.
Dal-Ré, R. (2018). Could phase 3 medicine trials be tagged as pragmatic? A case study: The Salford COPD trial. *Journal of evaluation in clinical practice, 24*, 258–261. doi:10.1111/jep.12796.
Daniel, G., Silcox, C., Bryan, J., McClellan, M., Romine, M., & frank, K. (2018). *Characterizing RWD quality and relevancy for regulatory purposes*. https://healthpolicy.duke.edu/sites/default/files/2020-08/Characterizing%20RWD%20for%20Regulatory%20Use.pdf.
Davis, C., Lexchin, J., Jefferson, T., Gøtzsche, P., & McKee, M. (2016). 'Adaptive pathways' to drug authorisation: adapting to industry? *BMJ, 354*, i4437. doi:10.1136/bmj.i4437.
Deaton, A., & Cartwright, N. (2018). Understanding and misunderstanding randomized controlled trials. *Social science & medicine (1982), 210*, 2–21. doi:10.1016/j.socscimed.2017.12.005.
Douglas, H. (2000). Inductive risk and values in science. *Philosophy of Science, 67*(4), 559–579.
Douglas, H. E. (2009). *Science, policy, and the value-free ideal*. University of Pittsburgh Press.
Downey, A., Wagner Gee, A., & Claiborne, A. B. (2017). *Real-world evidence generation and evaluation of therapeutics: Proceedings of a workshop*. Washington (DC). https://www.ncbi.nlm.nih.gov/books/NBK441700/.
Eichler, H.-G., Abadie, E., Breckenridge, A., Flamion, B., Gustafsson, L. L., Leufkens, H., et al. (2011). Bridging the efficacy-effectiveness gap: a regulator's perspective on addressing variability of drug response. *Nature reviews. Drug discovery, 10*, 495–506. doi:10.1038/nrd3501.
Eichler, H.-G., Bloechl-Daum, B., Brasseur, D., Breckenridge, A., Leufkens, H., Raine, J., et al. (2013). The risks of risk aversion in drug regulation. *Nature Reviews Drug Discovery, 12*, 907–916. doi:10.1038/nrd4129.
Eichler, H.-G., Bloechl-Daum, B., Broich, K., Kyrle, P. A., Oderkirk, J., Rasi, G., et al. (2019). Data rich, information poor: Can we use electronic health records to create a learning healthcare system for pharmaceuticals? *Clinical Pharmacology & Therapeutics, 105*, 912–922. doi:10.1002/cpt.1226.
Eichler, H.-G., Pignatti, F., Schwarzer-Daum, B., Hidalgo-Simon, A., Eichler, I., Arlett, P., et al. (2021). Randomized controlled trials versus real world evidence: neither magic nor myth. *Clinical Pharmacology & Therapeutics, 109*, 1212–1218. doi:10.1002/cpt.2083.
Embury, S. M., & Missier, P. (2014). Forget dimensions: Define your information quality using quality view patterns. In L. Floridi & P. Illari (Eds.), *The philosophy of information quality* (Synthese Library, Studies in Epistemology, Logic, Methodology, and Philosophy of Science, Vol. 358). Springer International Publishing.
Epstein, S. (1996). *Impure science: AIDS, activism, and the politics of knowledge* (Vol. 7). Berkeley, Calif.: University of California Press.
Erdman, J., Wolfram, J., Nimke, D., Croy, R., Wang, X., Weaver, T., et al. (2022). Lung transplant outcomes in adults in the united states: Retrospective cohort study using real-world evidence from the SRTR. *Transplantation, 106*, 1233–1242. doi:10.1097/TP.0000000000004011.
Ernst, E. (2002). What's the point of rigorous research on complementary/alternative medicine? *Journal of the Royal Society of Medicine, 95*, 211–213. doi:10.1258/jrsm.95.4.211.

European Medicines Agency. (2013). *Assessment report Relvar Ellipta* (Procedure No. EMEA/H/C/002673/0000). https://www.ema.europa.eu/en/documents/assessment-report/relvar-ellipta-epar-public-assessment-report_en.pdf. Accessed 26 November 2023.

European Medicines Agency. (2014). *Pilot project on adaptive licensing* (EMA/254350/2012). https://www.ema.europa.eu/system/files/documents/other/wc500163409_en.pdf. Accessed 5 January 2024.

European Medicines Agency. (2016). *Final report on the adaptive pathways pilot* (EMA/276376/2016). https://www.ema.europa.eu/system/files/documents/report/wc500211526_en.pdf. Accessed 3 January 2024.

European Medicines Agency. (2022). *Data quality framework for EU medicines regulation*. https://www.ema.europa.eu/system/files/documents/regulatory-procedural-guideline/data-quality-framework-eu-medicines-regulation_en_1.pdf. Accessed 7 February 2024.

European Medicines Agency. (2024). Big data. https://www.ema.europa.eu/en/about-us/how-we-work/big-data. Accessed 11 February 2024.

Faden, R. R., Beauchamp, T. L., & Kass, N. E. (2014). Informed consent, comparative effectiveness, and learning health care. *The New England journal of medicine, 370*, 766–768. doi:10.1056/NEJMhle1313674.

Farmer, R., Mathur, R., Bhaskaran, K., Eastwood, S. V., Chaturvedi, N., & Smeeth, L. (2018). Promises and pitfalls of electronic health record analysis. *Diabetologia, 61*, 1241–1248. doi:10.1007/s00125-017-4518-6.

Fischl, M. A., Richman, D. D., Grieco, M. H., Gottlieb, M. S., Volberding, P. A., Laskin, O. L., et al. (1987). The efficacy of azidothymidine (AZT) in the treatment of patients with AIDS and AIDS-related complex. A double-blind, placebo-controlled trial. *The New England Journal of Medicine (NEJM), 317*, 185–191. doi:10.1056/NEJM198707233170401.

Fleming, T. R., & DeMets, D. L. (1996). Surrogate end points in clinical trials: are we being misled? *Annals of internal medicine, 125*, 605–613. doi:10.7326/0003-4819-125-7-199610010-00011.

Floridi, L., & Illari, P. (Eds.). (2014). *The philosophy of information quality* (Synthese Library, Studies in Epistemology, Logic, Methodology, and Philosophy of Science, Vol. 358). Springer International Publishing.

Flynn, R., Plueschke, K., Quinten, C., Strassmann, V., Duijnhoven, R. G., Gordillo-Marañon, M., et al. (2022). Marketing authorization applications made to the european medicines agency in 2018-2019: What was the contribution of real-world evidence? *Clinical Pharmacology & Therapeutics, 111*, 90–97. doi:10.1002/cpt.2461.

Fraile Navarro, D., Tempini, N., & Teira, D. (2021). The trade-off between impartiality and freedom in the 21st Century Cures Act. *Philosophy of Medicine*. doi:10.5195/philmed.2021.24.

Franklin, J. M., Glynn, R. J., Martin, D., & Schneeweiss, S. (2019). Evaluating the use of nonrandomized real-world data analyses for regulatory decision making. *Clinical Pharmacology & Therapeutics, 105*, 867–877. doi:10.1002/cpt.1351.

Franklin, J. M., Glynn, R. J., Suissa, S., & Schneeweiss, S. (2020a). Emulation differences vs. biases when calibrating real-world evidence findings against randomized controlled trials. *Clinical Pharmacology & Therapeutics, 107*, 735–737. doi:10.1002/cpt.1793.

Franklin, J. M., Patorno, E., Desai, R. J., Glynn, R. J., Martin, D., Quinto, K., et al. (2021). Emulating randomized clinical trials with nonrandomized real-world evidence studies: first results from the RCT DUPLICATE initiative. *Circulation, 143*, 1002–1013. doi:10.1161/CIRCULATIONAHA.120.051718.

Franklin, J. M., Pawar, A., Martin, D., Glynn, R. J., Levenson, M., Temple, R., et al. (2020b). Nonrandomized real-world evidence to support regulatory decision making: Process for a randomized trial replication project. *Clinical Pharmacology & Therapeutics, 107*, 817–826. doi:10.1002/cpt.1633.

Franklin, J. M., & Schneeweiss, S. (2017). When and how can real world data analyses substitute for randomized controlled trials? *Clinical pharmacology and therapeutics, 102*, 924–933. doi:10.1002/cpt.857.

Fuller, J. (2013). Rationality and the generalization of randomized controlled trial evidence. *Journal of evaluation in clinical practice, 19*, 644–647. doi:10.1111/jep.12021.

Fuller, J. (2018). The confounding question of confounding causes in randomized trials. *The British Journal for the Philosophy of Science, 70*, 1–26. doi:10.1093/bjps/axx015.

Fuller, J. (2019). The myth and fallacy of simple extrapolation in medicine. *Synthese, 15*, 15. doi:10.1007/s11229-019-02255-0.

Funning, S., Grahnén, A., Eriksson, K., & Kettis-Linblad, Å. (2009). Quality assurance within the scope of Good Clinical Practice (GCP)—what is the cost of GCP-related activities? A survey within the Swedish Association of the Pharmaceutical Industry (LIF)'s members. *The Quality Assurance Journal, 12*, 3–7. doi:10.1002/qaj.433.

García-Albéniz, X., Hsu, J., & Hernán, M. A. (2017). The value of explicitly emulating a target trial when using real world evidence: an application to colorectal cancer screening. *European journal of epidemiology, 32*, 495–500. doi:10.1007/s10654-017-0287-2.

Gedeborg, R., Cline, C., Zethelius, B., & Salmonson, T. (2019). Pragmatic clinical trials in the context of regulation of medicines. *Upsala journal of medical sciences, 124*, 37–41. doi:10.1080/03009734.2018.1515280.

Gerber, A. S., & Green, D. P. (2012). *Field experiments: Design, analysis, and interpretation*. W.W. Norton & Company.

Glaettli, M., Di Petto, L., & Fayet-Mello, A. (2022). Clinical research core competencies framework. https://smw.ch/index.php/smw/announcement/view/55. Accessed 10.02.24.

Gloy, V., Schmitt, A. M., Düblin, P., Hirt, J., Axfors, C., Kuk, H., et al. (2023). The evidence base of US Food and Drug Administration approvals of novel cancer therapies from 2000 to 2020. *International journal of cancer, 152*, 2474–2484. doi:10.1002/ijc.34473.

Godwin, M., Ruhland, L., Casson, I., MacDonald, S., Delva, D., Birtwhistle, R., et al. (2003). Pragmatic controlled clinical trials in primary care: the struggle between external and internal validity. *BMC Medical Research Methodology, 3*, 28. doi:10.1186/1471-2288-3-28.

Green, S., Hillersdal, L., Holt, J., Hoeyer, K., & Wadmann, S. (2022). The practical ethics of repurposing health data: how to acknowledge invisible data work and the need for prioritization. *Medicine, health care, and philosophy*, 1–14. doi:10.1007/s11019-022-10128-6.

Grey, A., Bolland, M. J., Avenell, A., Klein, A. A., & Gunsalus, C. K. (2020). Check for publication integrity before misconduct. *Nature, 577*, 167–169. doi:10.1038/d41586-019-03959-6.

Guala, F. (2003). Experimental localism and external validity. *Philosophy of Science, 70*, 1195–1205. doi:10.1086/377400.

Guala, F. (2010). Extrapolation, analogy, and comparative process tracing. *Philosophy of Science, 77*, 1070–1082. doi:10.1086/656541.

H. Chytilová, & R. Maialeh. (2015). Internal and external validity in experimental economics. *World Academy of Science, Engineering and Technology, International Journal of Social, Behavioral, Educational, Economic, Business and Industrial Engineering, 9*(6).

Hansen, K., & Kappel, K. (2010). The proper role of evidence in complementary/alternative medicine. *Journal of Medicine and Philosophy, 35*, 7–18. doi:10.1093/jmp/jhp059.

Hatswell, A. J., Baio, G., Berlin, J. A., Irs, A., & Freemantle, N. (2016). Regulatory approval of pharmaceuticals without a randomised controlled study: analysis of EMA and FDA approvals 1999-2014. *BMJ open, 6*, 1-8. doi:10.1136/bmjopen-2016-011666.

Hauray, B. (2017). From regulatory knowledge to regulatory decisions: The european evaluation of medicines. *Minerva, 55*, 187–208. doi:10.1007/s11024-017-9323-3.

Hemkens, L. G. (2018). How routinely collected data for randomized trials provide long-term randomized real-world evidence. *JAMA network open, 1*, 1-3. doi:10.1001/jamanetworkopen.2018.6014.

Hemkens, L. G., Contopoulos-Ioannidis, D. G., & Ioannidis, J. P. A. (2016). Routinely collected data and comparative effectiveness evidence: promises and limitations. *CMAJ : Canadian Medical Association Journal, 188*, 158-164. doi:10.1503/cmaj.150653.

Herper, M. (2019). Attempt to replicate clinical trials with real-world data generates real-world criticism, too. Boston Globe Life Sciences Media, LLC. https://www.statnews.com/2019/07/03/replicate-clinical-trials-real-world-evidence/. Accessed 12.2.24.

Herrett, E., Gallagher, A. M., Bhaskaran, K., Forbes, H., Mathur, R., van Staa, T., et al. (2015). Data resource profile: Clinical practice research datalink (CPRD). *International Journal of Epidemiology, 44*, 827–836. doi:10.1093/ije/dyv098.

Higgins, J. P. T., Savovic, J., Page, M. J., & Sterne, J. A. C. (2019). *RoB 2: a revised tool for assessing risk of bias in randomised trials.*. https://www.riskofbias.info/welcome/rob-2-0-tool/current-version-of-rob-2.

Hirt, J., Janiaud, P., Düblin, P., & Hemkens, L. G. (2023). Meta-research on pragmatism of randomized trials: rationale and design of the PragMeta database. *Trials, 24*, 1–8. doi:10.1186/s13063-023-07474-y.

Hirt, J., Janiaud, P., Düblin, P., Nicoletti, G. J., Dembowska, K., Nguyen, T. V. T., et al. (2024). Use of pragmatic randomized trials in multiple sclerosis: A systematic overview. *Multiple sclerosis (Houndmills, Basingstoke, England)*, 13524585231221938. doi:10.1177/13524585231221938.

Howick, J. (2011). *The philosophy of evidence-based medicine*. Oxford, UK: Wiley-Blackwell.

Howick, J. (2017). The relativity of 'placebos': defending a modified version of Grünbaum's definition. *Synthese, 194*, 1363–1396. doi:10.1007/s11229-015-1001-0.

Howick, J., Glasziou, P., & Aronson, J. K. (2013a). Can understanding mechanisms solve the problem of extrapolating from study to target populations (the problem of 'external validity')? *Journal of the Royal Society of Medicine, 106*, 81–86. doi:10.1177/0141076813476498.

Howick, J., Glasziou, P., & Aronson, J. K. (2013b). Problems with using mechanisms to solve the problem of extrapolation. *Theoretical medicine and bioethics, 34*, 275–291. doi:10.1007/s11017-013-9266-0.

Howson, C., & Urbach, P. (2006). *Scientific reasoning: The Bayesian approach* (3rd ed.). Chicago: Open Court.

Hutchinson, N., Moyer, H., Zarin, D. A., & Kimmelman, J. (2022). The proportion of randomized controlled trials that inform clinical practice. *eLife*. doi:10.7554/eLife.79491.

Illari, P. (2014). IQ: Purpose and dimensions. In L. Floridi & P. Illari (Eds.), *The philosophy of information quality* (Vol. 358, pp. 281–301, Synthese Library, Studies in Epistemology, Logic, Methodology, and Philosophy of Science, Vol. 358). Springer International Publishing.

International Council for Harmonisation. (1998). *Statistical principles for clinical trials E9*. https://database.ich.org/sites/default/files/E9_Guideline.pdf. Accessed 7 February 2024.
International Council for Harmonisation. (2016). *Guideline for Good Clinical Practice (GCP) E6-R2: Integrated addendum to ICH E6-R1* .
International Council for Harmonisation. (2021a). *Overview of ICH*. https://admin.ich.org/sites/default/files/2021-06/OverviewOfICH_2021_0614.pdf. Accessed 10.12.23.
International Council for Harmonisation. (2021b). *ICH guideline E8 (R1) on general considerations for clinical studies* (ICH-E8 (R1)).
International Council for Harmonisation. (2023). *Good Clinical Practice (GCP) E6-R3: Draft version, under public consultation*. https://www.ema.europa.eu/en/ich-e6-r2-good-clinical-practice-scientific-guideline#ema-inpage-item-8264. Accessed 28 July 2023.
Inverso, G., Flath-Sporn, S. J., Monoxelos, L., Labow, B. I., Padwa, B. L., & Resnick, C. M. (2016). What is the cost of meaningful use? *Journal of oral and maxillofacial surgery : official journal of the American Association of Oral and Maxillofacial Surgeons, 74*, 227–229. doi:10.1016/j.joms.2015.10.010.
Ioannidis, J. P. A. (2016). Why most clinical research Is not useful. *PLoS Medicine, 13*, e1002049. doi:10.1371/journal.pmed.1002049.
Irving, E., van den Bor, R., Welsing, P., Walsh, V., Alfonso-Cristancho, R., Harvey, C., et al. (2017). Series: Pragmatic trials and real world evidence: Paper 7. Safety, quality and monitoring. *Journal of clinical epidemiology, 91*, 6–12. doi:10.1016/j.jclinepi.2017.05.004.
Irzik, G., & Kurtulmus, F. (2019). What is epistemic public trust in science? *The British Journal for the Philosophy of Science, 70*, 1145–1166. doi:10.1093/bjps/axy007.
James, S., Rao, S. V., & Granger, C. B. (2015). Registry-based randomized clinical trials-a new clinical trial paradigm. *Nature reviews. Cardiology, 12*, 312–316. doi:10.1038/nrcardio.2015.33.
Jimenez-Buedo, M., & Miller, L. M. (2010). Why a trade-off? The relationship between the external and internal validity of experiments. *Theoria : An International Journal for Theory, History and Fundations of Science, 25*, 301–321. doi:10.1387/theoria.779.
John, S. (2018). Epistemic trust and the ethics of science communication: against transparency, openness, sincerity and honesty. *Social Epistemology, 32*, 75–87. doi:10.1080/02691728.2017.1410864.
John, S. (2021). Science, politics and regulation: The trust-based approach to the demarcation problem. *Studies in history and philosophy of science, 90*, 1–9. doi:10.1016/j.shpsa.2021.08.006.
Jonker, C. J., Bakker, E., Kurz, X., & Plueschke, K. (2022). Contribution of patient registries to regulatory decision making on rare diseases medicinal products in Europe. *Frontiers in Pharmacology, 13*, 1–9. doi:10.3389/fphar.2022.924648.
Kahn, M. G., Callahan, T. J., Barnard, J., Bauck, A. E., Brown, J., Davidson, B. N., et al. (2016). A harmonized data quality assessment terminology and framework for the secondary use of electronic health record data. *eGEMs, 4*, 1244. doi:10.13063/2327-9214.1244.
Kahn, M. G., Eliason, B. B., & Bathurst, J. (2010). Quantifying clinical data quality using relative gold standards. *AMIA Annual Symposium Proceedings, 2010*, 356–360.
Kalkman, S., van Thiel, G., van der Graaf, R., Zuidgeest, M., Goetz, I., Grobbee, D., et al. (2017a). The social value of pragmatic trials. *Bioethics, 31*, 136–143. doi:10.1111/bioe.12315.
Kalkman, S., van Thiel, G. J. M. W., Zuidgeest, M. G. P., Goetz, I., Pfeiffer, B. M., Grobbee, D. E., et al. (2017b). Series: Pragmatic trials and real world evidence:

Paper 4. Informed consent. *Journal of clinical epidemiology, 89*, 181–187. doi:10.1016/j.jclinepi.2017.03.019.
Karanicolas, P. J., Montori, V. M., Devereaux, P. J., Schünemann, H., & Guyatt, G. H. (2009). A new 'mechanistic-practical" framework for designing and interpreting randomized trials. *Journal of clinical epidemiology, 62*, 479–484. doi:10.1016/j.jclinepi.2008.02.009.
Kombe, M. M., Zulu, J. M., Michelo, C., & Sandøy, I. F. (2019). Community perspectives on randomisation and fairness in a cluster randomised controlled trial in Zambia. *BMC medical ethics, 20*, 1–10. doi:10.1186/s12910-019-0421-7.
Kwakkenbos, L., Imran, M., McCall, S. J., McCord, K. A., Fröbert, O., Hemkens, L. G., et al. (2021). CONSORT extension for the reporting of randomised controlled trials conducted using cohorts and routinely collected data (CONSORT-ROUTINE): checklist with explanation and elaboration. *BMJ, 373*, n857. doi:10.1136/bmj.n857.
La Caze, A. (2017). The randomized controlled trial: internal and external validity. In *The Routledge companion to philosophy of medicine* (pp. 195–206, Routledge Philosophy Companions). Milton.
Landes, J., Osimani, B., & Poellinger, R. (2018). Epistemology of causal inference in pharmacology. *European Journal for Philosophy of Science, 8*, 3–49. doi:10.1007/s13194-017-0169-1.
Lareau, S. C., & Yawn, B. P. (2010). Improving adherence with inhaler therapy in COPD. *International Journal of Chronic Obstructive Pulmonary Disease, 5*, 401–406. doi:10.2147/COPD.S14715.
Larroulet Philippi, C. (2022). There is cause to randomize. *Philosophy of Science, 89*, 152–170. doi:10.1017/psa.2021.19.
Leonelli, S. (2016). *Data-centric biology: A philosophical study*. Chicago: The University of Chicago Press.
Leonelli, S. (2017a). *Biomedical knowledge production in the age of big data: Analysis conducted on behalf of the Swiss Science and Innovation Council SSIC*. https://wissenschaftsrat.ch/images/stories/pdf/en/Exploratory_study_2_2017_Big_Data_SSIC_EN.pdf.
Leonelli, S. (2017b). Global data quality assessment and the situated nature of 'best' research practices in biology. *1683-1470*. doi:10.5334/dsj-2017-032.
Leonelli, S. (2020). Learning from data journeys. In S. Leonelli & N. Tempini (Eds.), *Data journeys in the sciences* (1-27). Cham: Springer International Publishing.
Leonelli, S., & Tempini, N. (Eds.). (2020). *Data journeys in the sciences*. Cham: Springer International Publishing.
Leppke, S., Leighton, T., Zaun, D., Chen, S.-C., Skeans, M., Israni, A. K., et al. (2013). Scientific registry of transplant recipients: collecting, analyzing, and reporting data on transplantation in the united states. *Transplantation reviews, 27*, 50–56. doi:10.1016/j.trre.2013.01.002.
Loudon, K., Treweek, S., Sullivan, F., Donnan, P., Thorpe, K. E., & Zwarenstein, M. (2015). The PRECIS-2 tool: designing trials that are fit for purpose. *BMJ (Clinical research ed.), 350*, 1-11. doi:10.1136/bmj.h2147.
Luce, B. R., Drummond, M., Jönsson, B., Neumann, P. J., Schwartz, J. S., Siebert, U., et al. (2010). EBM, HTA, and CER: clearing the confusion. *The Milbank quarterly, 88*, 256–276. doi:10.1111/j.1468-0009.2010.00598.x.
Mahendraratnam, N., Mercon, K., Gill, M., Benzing, L., & McClellan, M. B. (2022). Understanding use of real-world data and real-world evidence to support regulatory decisions on medical product effectiveness. *Clinical Pharmacology & Therapeutics, 111*, 150–154. doi:10.1002/cpt.2272.
Martinez, M., & Teira, D. (2021). Why experimental balance is still a reason to randomize. *The British Journal for the Philosophy of Science*. doi:10.1086/716096.

Mc Cord, K. A., Al-Shahi Salman, R., Treweek, S., Gardner, H., Strech, D., Whiteley, W., et al. (2018). Routinely collected data for randomized trials: promises, barriers, and implications. *Trials, 19*, 29. doi:10.1186/s13063-017-2394-5.

Mc Cord, K. A., & Hemkens, L. G. (2019). Using electronic health records for clinical trials: Where do we stand and where can we go? *CMAJ : Canadian Medical Association Journal, 191*, E128-33. doi:10.1503/cmaj.180841.

Mehra, M. R., Desai, S. S., Ruschitzka, F., & Patel, A. N. (2020). RETRACTED: Hydroxychloroquine or chloroquine with or without a macrolide for treatment of COVID-19: a multinational registry analysis. *The Lancet*. doi:10.1016/S0140-6736(20)31180-6.

Monitoring Platform of the SCTO. (2023). *Position paper: Monitoring in non-interventional human research projects*. Swiss Clinical Trial Organisation (SCTO).

Munos, B. (2009). Lessons from 60 years of pharmaceutical innovation. *Nature Reviews Drug Discovery, 8*, 959–968. doi:10.1038/nrd2961.

Nathan, R. A., Sorkness, C. A., Kosinski, M., Schatz, M., Li, J. T., Marcus, P., et al. (2004). Development of the asthma control test: a survey for assessing asthma control. *The Journal of allergy and clinical immunology, 113*, 59–65. doi:10.1016/j.jaci.2003.09.008.

NIH Pragmatic Trials Collaboratory. (2014). *Assessing data quality for healthcare systems data used in clinical research*. https://dcricollab.dcri.duke.edu/sites/NIHKR/KR/Assessing-data-quality.pdf#search=data%20quality. Accessed 8 February 2024.

NIH Pragmatic Trials Collaboratory. (2024). *The living textbook: Asessing fitness for use of real-world data sources*. https://rethinkingclinicaltrials.org/chapters/conduct/assessing-fitness-for-use-of-real-world-data-sources/introduction/. Accessed 8 February 2024.

Nordon, C., Karcher, H., Groenwold, R. H., Ankarfeldt, M. Z., Pichler, F., Chevrou-Severac, H., et al. (2016). The "efficacy-effectiveness gap": Historical background and current conceptualization. *Value in health, 19*, 75–81. doi:10.1016/j.jval.2015.09.2938.

Offord, C. (2020). The Surgisphere Scandal: What Went Wrong? Labx media gropu. https://www.the-scientist.com/features/the-surgisphere-scandal-what-went-wrong--67955. Accessed 9 December 2023.

Organ Procurement and Transplantation Network (OPTN). (2023). *OPTN Policies: Effective Date 12/13/2023*. https://optn.transplant.hrsa.gov/media/eavh5bf3/optn_policies.pdf. Accessed 8 February 2024.

Osimani, B. (2013). Hunting side effects and explaining them: Should we reverse evidence hierarchies upside down? *Topoi, 33*, 295–312. doi:10.1007/s11245-013-9194-7.

Patino, C. M., & Ferreira, J. C. (2018). Internal and external validity: can you apply research study results to your patients? *Jornal brasileiro de pneumologia : publicacao oficial da Sociedade Brasileira de Pneumologia e Tisilogia, 44*, 183. doi:10.1590/S1806-37562018000000164.

Patsopoulos, N. A. (2011). A pragmatic view on pragmatic trials. *Dialogues in Clinical Neuroscience, 13*, 217–224. doi:10.31887/DCNS.2011.13.2/npatsopoulos.

Persson, J., & Wallin, A. (2015). The (misconceived) distinction between internal and external validity. In E. Sjöstrand, J. Persson, G. Hermerén, & N.-E. Sahlin (Eds.), *Against boredom: 17 essays : on ignorance, values, creativity, metaphysics, decision-making, truth, preference, art, processes, Ramsey, ethics, rationality, validity, human ills, science and eternal life : to Nils-Eric Sahlin on the occasion of his 60th birthday* (pp. 87–195). Lidingö: Fri Tanke.

Pietsch, W. (2021). *Big data* (Cambridge elements. Elements in the philosophy of science). Cambridge: Cambridge University Press.
Pirosca, S., Shiely, F., Clarke, M., & Treweek, S. (2022). Tolerating bad health research: the continuing scandal. *Trials, 23*, 458. doi:10.1186/s13063-022-06415-5.
PSI RWD SIG. (2021). Webinar: Real-world evidence submission - a case study in lung transplantation. https://psiweb.org/vod/item/psi-rwd-sig-webinar-real-world-evidence-submission---a-case-study-in-lung-transplantation. Accessed 20.03.23.
Quint, J. K., Müllerova, H., DiSantostefano, R. L., Forbes, H., Eaton, S., Hurst, J. R., et al. (2014). Validation of chronic obstructive pulmonary disease recording in the Clinical Practice Research Datalink (CPRD-GOLD). *BMJ open, 4*, 1-8. doi:10.1136/bmjopen-2014-005540.
Rawlins, M. D. (2004). Cutting the cost of drug development? *Nature Reviews Drug Discovery, 3*, 360–364. doi:10.1038/nrd1347.
Reiss, J. (2019). Against external validity. *Synthese, 196*, 3103–3121. doi:10.1007/s11229-018-1796-6.
Rocca, E., & Anjum, R. L. (2020). Causal evidence and dispositions in medicine and public health. *International Journal of Environmental Research and Public Health, 17*, 1–18. doi:10.3390/ijerph17061813.
Rothnie, K. J., Müllerová, H., Hurst, J. R., Smeeth, L., Davis, K., Thomas, S. L., et al. (2016). Validation of the recording of acute exacerbations of copd in uk primary care electronic healthcare records. *PLoS ONE, 11*, 1-14. doi:10.1371/journal.pone.0151357.
Rothwell, P. M. (2005). External validity of randomised controlled trials: "To whom do the results of this trial apply?". *The Lancet, 365*, 82–93. doi:10.1016/S0140-6736(04)17670-8.
Roy, A. S. A. (2012). *Stifling new cures: the true cost of lengthy clinical drug trials* (Project FDA Report 05). https://manhattan.institute/article/stifling-new-cures-the-true-cost-of-lengthy-clinical-drug-trials. Accessed 30 December 2023.
Sackett, D. L. (2007). Commentary: Measuring the success of blinding in RCTs: don't, must, can't or needn't? *International Journal of Epidemiology, 36*, 664–665. doi:10.1093/ije/dym088.
Schneeweiss, S. (2016). Improving therapeutic effectiveness and safety through big healthcare data. *Clinical Pharmacology & Therapeutics, 99*, 262–265. doi:10.1002/cpt.316.
Schneeweiss, S. (2019). Real-world evidence of treatment effects: The useful and the misleading. *Clinical Pharmacology & Therapeutics, 106*, 43–44. doi:10.1002/cpt.1405.
Schneeweiss, S., & Avorn, J. (2005). A review of uses of health care utilization databases for epidemiologic research on therapeutics. *Journal of clinical epidemiology, 58*, 323–337. doi:10.1016/j.jclinepi.2004.10.012.
Schneidermann, M. (1966). Therapeutic trials in cancer: Working paper prepared for WHO Expert Committee on Cancer Treatment, Geneva, Switzerland, 9-15 March 1965. WHO/CANC/66.66. https://www.jameslindlibrary.org/schneidermann-ma-1966/. Accessed 26 November 2023.
Schram, A. (2005). Artificiality: The tension between internal and external validity in economic experiments. *Journal of Economic Methodology, 12*, 225–237. doi:10.1080/13501780500086081.
Schulz, K. F., Chalmers, I., & Altman, D. G. (2002). The landscape and lexicon of blinding in randomized trials. *Annals of internal medicine, 136*, 254–259. doi:10.7326/0003-4819-136-3-200202050-00022.
Schwartz, D., & Lellouch, J. (2009). Explanatory and pragmatic attitudes in therapeutical trials. *Journal of clinical epidemiology, 62*, 499–505. doi:10.1016/j.jclinepi.2009.01.012.

Senn, S. (1989). Covariate imbalance and random allocation in clinical trials. *Statistics in Medicine, 8*, 467–475. doi:10.1002/sim.4780080410.
Senn, S. (1994). Fisher's game with the devil. *Statistics in medicine, 13*, 217–230. doi:10.1002/sim.4780130305.
Senn, S. (2013). Seven myths of randomisation in clinical trials. *Statistics in medicine, 32*, 1439–1450. doi:10.1002/sim.5713.
Senn, S. (2020). Randomisation is not about balance, nor about homogeneity but about randomness (Guest Post). https://errorstatistics.com/2020/04/20/s-senn-randomisation-is-not-about-balance-nor-about-homogeneity-but-about-randomness-guest-post/. Accessed 20 September 2021.
Simon, G. E., Shortreed, S. M., Rossom, R. C., Penfold, R. B., Sperl-Hillen, J. A. M., & O'Connor, P. (2019). Principles and procedures for data and safety monitoring in pragmatic clinical trials. *Trials, 20*, 1–8. doi:10.1186/s13063-019-3869-3.
Smith, R. (2014). *Medical research—still a scandal*. The BMJ Opinion. https://blogs.bmj.com/bmj/2014/01/31/richard-smith-medical-research-still-a-scandal/. Accessed 5 January 2023.
Solomon, M. (2015). *Making medical knowledge*. Oxford University Press.
Steel, D. (2008). *Across the boundaries: Extrapolation in biology and social science* (Environmental ethics and science policy series). Oxford University Press.
Stegenga, J. (2014). Information quality in clinical research. In L. Floridi & P. Illari (Eds.), *The philosophy of information quality* (Vol. 358, pp. 163–182, Synthese Library, Studies in Epistemology, Logic, Methodology, and Philosophy of Science, Vol. 358). Springer International Publishing.
Stegenga, J. (2017). Drug regulation and the inductive risk calculus. In K. C. Elliott & T. Richards (Eds.), *Exploring inductive risk: Case studies of values in science* (pp. 17–36, Vol. 1). Oxford University Press.
Stegenga, J. (2018). *Medical nihilism*. Oxford University Press.
Stock, P. G. (2017). Strengths and weaknesses of using SRTR data to shape the management of the HIV-infected kidney transplant recipient. *American journal of transplantation : official journal of the American Society of Transplantation and the American Society of Transplant Surgeons, 17*, 3001–3002. doi:10.1111/ajt.14479.
Swissmedic. (2022). *Swissmedic position paper on the use of real world evidence*. https://www.swissmedic.ch/swissmedic/en/home/news/mitteilungen/positionspapier-verwendung-real-world-evidence.html. Accessed 30 September 2022.
Teira, D. (2013). Blinding and the non-interference assumption in medical and social trials. *Philosophy of the Social Sciences, 43*, 358–372. doi:10.1177/0048393113488871.
Teira, D. (2016). Debiasing methods and the acceptability of experimental outcomes. *Perspectives on Science, 24*, 722–743. doi:10.1162/POSC_a_00230.
Teira, D. (2020). On the normative foundations of pharmaceutical regulation. In A. LaCaze (Ed.), *Uncertainty in Pharmacology* (Vol. 338, pp. 417–437, Boston Studies in the Philosophy and History of Science). Cham: Springer International Publishing.
Teira, D., & Reiss, J. (2013). Causality, impartiality and evidence-based policy. In H.-K. Chao, S.-T. Chen, & R. L. Millstein (Eds.), *Mechanism and causality in biology and economics* (Vol. 3, pp. 207–224, History, Philosophy and Theory of the Life Sciences, Vol. 3). Dordrecht, s.l.: Springer Netherlands.
Tempini, N., & Leonelli, S. (2018). Concealment and discovery: The role of information security in biomedical data re-use. *Social Studies of Science, 48*, 663–690. doi:10.1177/0306312718804875.
Tempini, N., & Teira, D. (2020). The babel of drugs: on the consequences of evidential pluralism in pharmaceutical regulation and regulatory data journeys. In

S. Leonelli & N. Tempini (Eds.), *Data journeys in the sciences* (pp. 207–225). Cham: Springer International Publishing.
Thorpe, K. E., Zwarenstein, M., Oxman, A. D., Treweek, S., Furberg, C. D., Altman, D. G., et al. (2009). A pragmatic-explanatory continuum indicator summary (PRECIS): a tool to help trial designers. *Journal of clinical epidemiology, 62*, 464–475. doi:10.1016/j.jclinepi.2008.12.011.
Tonelli, M. R., & Callahan, T. C. (2001). Why alternative medicine cannot be evidence-based. *Academic medicine : journal of the Association of American Medical Colleges, 76*, 1213–1220. doi:10.1097/00001888-200112000-00011.
Travers, J., Marsh, S., Caldwell, B., Williams, M., Aldington, S., Weatherall, M., et al. (2007). External validity of randomized controlled trials in COPD. *Respiratory medicine, 101*, 1313–1320. doi:10.1016/j.rmed.2006.10.011.
Tripepi, G., Chesnaye, N. C., Dekker, F. W., Zoccali, C., & Jager, K. J. (2020). Intention to treat and per protocol analysis in clinical trials. *Nephrology (Carlton, Vic.), 25*, 513–517. doi:10.1111/nep.13709.
Tunis, S. R., Stryer, D. B., & Clancy, C. M. (2003). Practical clinical trials: increasing the value of clinical research for decision making in clinical and health policy. *JAMA, 290*, 1624–1632. doi:10.1001/jama.290.12.1624.
US Food and Drug Administration. (2018). Framework for FDA's real-world evidence program. Available online at: https://www.fda.gov/downloads/ScienceResearch/SpecialTopics/RealWorldEvidence/UCM627769.pdf. Accessed 19.01.19.
US Food and Drug Administration. (2021a). FDA approves new use of transplant drug based on real-world evidence. https://www.fda.gov/drugs/news-events-human-drugs/fda-approves-new-use-transplant-drug-based-real-world-evidence. Accessed 19 February 2023.
US Food and Drug Administration. (2023). Real-world evidence. https://www.fda.gov/science-research/science-and-research-special-topics/real-world-evidence. Accessed 14.12.23.
US Food and Drug Administration (CBER). (2021b). *Real-world data: Assessing electronic health records and medical claims data to support regulatory decisionmaking for drug and biological products guidance for industry draft guidance (for public consultation): Guidance for industry*. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/real-world-data-assessing-electronic-health-records-and-medical-claims-data-support-regulatory.
US Food and Drug Administration (CBER). (2021c). *Real-world data: Assessing registries to support regulatory decision-making for drug and biological products guidance for industry: Guidance for industry*. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/real-world-data-assessing-registries-support-regulatory-decision-making-drug-and-biological-products. Accessed 4 October 2022.
Vestbo, J., Leather, D., Diar Bakerly, N., New, J., Gibson, J. M., McCorkindale, S., et al. (2016). Effectiveness of fluticasone furoate-vilanterol for COPD in clinical practice. *The New England journal of medicine, 375*, 1253–1260. doi:10.1056/NEJMoa1608033.
Wallach, J. D., Zhang, A. D., Skydel, J. J., Bartlett, V. L., Dhruva, S. S., Shah, N. D., et al. (2021). Feasibility of using real-world data to emulate postapproval confirmatory clinical trials of therapeutic agents granted us food and drug administration accelerated approval. *JAMA network open, 4*, e2133667. doi:10.1001/jamanetworkopen.2021.33667.
Weiskopf, N. G., & Weng, C. (2013). Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. *Journal of the American Medical Informatics Association : JAMIA, 20*, 144–151. doi:10.1136/amiajnl-2011-000681.

Wilholt, T. (2013). Epistemic trust in Science. *The British Journal for the Philosophy of Science, 64*, 233–253. doi:10.1093/bjps/axs007.

Wilholt, T. (2016). Collaborative research, scientific communities, and the social diffusion of trustworthiness. In M. S. Brady & M. Fricker (Eds.), *The epistemic life of groups*: *Essays in the epistemology of collectives* (pp. 218–234, Mind Association occasional series). Oxford: Oxford University Press.

Woodcock, A., Boucot, I., Leather, D. A., Crawford, J., Collier, S., Bakerly, N. D., et al. (2018). Effectiveness versus efficacy trials in COPD: how study design influences outcomes and applicability. *European Respiratory Journal, 51*, 2–11. doi:10.1183/13993003.01531-2017.

Worrall, J. (2002). What evidence in evidence-based medicine? *Philosophy of Science, 69*(69), 316-330.

Worrall, J. (2007). Why there's no C^cause to randomize. *The British Journal for the Philosophy of Science, 58*, 451–488. doi:10.1093/bjps/axm024.

Wouters, O. J., McKee, M., & Luyten, J. (2020). Estimated Research and Development Investment Needed to Bring a New Medicine to Market, 2009-2018. *JAMA, 323*, 844–853. doi:10.1001/jama.2020.1166.

Yanik, E. L., Nogueira, L. M., Koch, L., Copeland, G., Lynch, C. F., Pawlish, K. S., et al. (2016). Comparison of cancer diagnoses between the us solid organ transplant registry and linked central cancer registries. *American journal of transplantation, 16*, 2986–2993. doi:10.1111/ajt.13818.

Zuidgeest, M. G. P., Goetz, I., Groenwold, R. H. H., Irving, E., van Thiel, G. J. M. W., & Grobbee, D. E. (2017). Series: Pragmatic trials and real world evidence: Paper 1. Introduction. *Journal of clinical epidemiology*, 7–13. doi:10.1016/j.jclinepi.2016.12.023.

Zwarenstein, M., & Treweek, S. (2009). What kind of randomized trials do we need? *CMAJ : Canadian Medical Association Journal, 180*, 998–1000. doi:10.1503/cmaj.082007.

Zwarenstein, M., Treweek, S., Gagnier, J. J., Altman, D. G., Tunis, S., Haynes, B., et al. (2008). Improving the reporting of pragmatic trials: an extension of the CONSORT statement. *BMJ, 337*, 1-8. doi:10.1136/bmj.a2390.

114th Congress (2015-2016). *H.R.34 - 21st Century Cures Act* .