International Studies in the Philosophy of Science

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/cisp20

Assessing the Overall Validity of Randomised Controlled Trials

Alexander Krauss

To cite this article: Alexander Krauss (2021): Assessing the Overall Validity of Randomised Controlled Trials, International Studies in the Philosophy of Science, DOI: 10.1080/02698595.2021.2002676

To link to this article: https://doi.org/10.1080/02698595.2021.2002676

INTERNATIONAL STUDIES IN THE PHILOSOPHY OF SCIENCE
https://doi.org/10.1080/02698595.2021.2002676

Routledge
Taylor & Francis Group

Assessing the Overall Validity of Randomised Controlled Trials

Alexander Krauss[a,b]

[a]London School of Economics; [b]University of Barcelona

ABSTRACT

In the biomedical, behavioural and social sciences, the leading method used to estimate causal effects is commonly randomised controlled trials (RCTs) that are generally viewed as both the source and justification of the most valid evidence. In studying the foundation and theory behind RCTs, the existing literature analyses important single issues and biases in isolation that influence causal outcomes in trials (such as randomisation, statistical probabilities and placebos). The common account of biased causal inference is described in a general way in terms of probabilistic imbalances between trial groups. This paper expands the common account of causal bias by distinguishing between the range of biases arising *between* trial groups but also *within* one of the groups or *across* the entire sample during trial design, implementation and analysis. This is done by providing concrete examples from highly influential RCT studies. In going beyond the existing RCT literature, the paper provides a broader, practice-based account of causal bias that specifies the between-group, within-group and across-group biases that affect the estimated causal results of trials – impacting both the effect size and statistical significance. Within this expanded framework, we can better identify the range of different types of biases we face in practice and address the central question about the overall validity of the RCT method and its causal claims. A study can face several smaller biases (related simultaneously to a smaller sample, smaller estimated effect, greater unblinding etc.) that generally add up to greater *aggregate bias*. Though difficult to measure precisely, it is important to assess and provide information in studies on how much different sources of bias, combined, can explain the estimated causal effect. The RCT method is thereby often the best we have to inform our policy decisions – and the evidence is strengthened when combined with multiple studies and other methods. Yet there is room for continually improving trials and identifying ways to reduce biases they face and to increase their overall validity. Implications are discussed.

KEYWORDS

Philosophy of science; philosophy of medicine; randomised controlled trials; RCTs; bias; validity; internal validity

CONTACT Alexander Krauss a.krauss@lse.ac.uk

Introduction

Many of us have likely used some medication, own some technology or supported some public policy tested in a trial. To be able to assess how effective they may be prior to supporting them – either as patients, consumers or voters – RCTs are often carried out by randomly splitting up a sample of people into either a treatment group (who receive the treatment or intervention) or a control group (who do not receive it). The RCT method has revolutionised the medical sciences in the second half of the twentieth century and many of the behavioural and social sciences since then. Across the medical sciences[1] and behavioural and social sciences,[2] RCTs are commonly viewed as the best means to generate causal knowledge about what works by providing an 'unbiased estimate of the average effect' of a treatment (see, Gorard and Taylor 2004, 94; Kane 2006, 118; Waters et al. 2010; Kelly, Lesh, and Baek 2014, 29).

Nancy Cartwright thereby outlines that RCTs are based on 'a deductive method: if the assumptions of the test are met, a positive result implies the appropriate causal conclusion' (Cartwright 2007, 11; cf. 2010). That is, 'positive results in an ideal RCT with treatment C and outcome E deductively implies C causes E in the experimental population' (Cartwright 2010, 68). This common view of the ideal RCT – that its causal conclusions can be derived deductively and validly from the evidence to 'secure internal validity' (Cartwright 2007, 2010) – can help guide the design of real RCTs, and help justify experimental choices. Experiments are however always conducted in the real world and fall short of the ideal. That is, the ideal RCT does not always reflect the complex practice of conducting trials. Causal conclusions in trials are, in practice, inferred from estimated statistical results that are the outcome of multiple complex processes including selecting a sample, randomising, blinding, controlling, carrying out treatments, monitoring the level of adherence among participants, etc. And these processes involve many actors making many decisions at different steps when designing, implementing and analysing trials. The common view of causal bias in the RCT literature moreover captures one way in which bias can arise in trials, namely via probabilistic imbalances between trial groups (Cartwright 2007, 2010). Causal bias however not only arises *between* groups especially during trial design – but it also arises *within* one of the groups, or *across* the entire sample itself during trial design, implementation and analysis that also influences the estimated causal results. In going beyond and expanding the common view of the RCT method (cf. Cartwright 2007, 2010), the paper provides a practice-based account of trials that illustrates – using empirical evidence from the ten most cited RCT studies – that:

(i) the complex processes involved in trials bring important assumptions that cannot generally all be fully met in practice,
(ii) the assumptions and biases can constrain the overall validity of trials, and
(iii) a broader view of causal bias in trials is needed that enables us to better identify the range of possible biases (and thus better improve trials) by incorporating biases arising not only between trial groups but also within a group or across the overall sample that are as important in influencing a trial's causal outcomes and overall validity.

Other philosophers, beyond Cartwright, have also studied the foundation and theory behind the RCT method, and its validity. Worrall (2007, 2007a, 2010) assesses the function and limitation of randomisation, and reasons about probabilistic causality. Howick (2017, 2011) analyses the role of the placebo in affecting trial results. Holman and Bruner (2017) argue that a funder bias can arise through individual researchers and an industry-funded community that can affect causal claims in trials. Reiss (2019), Favereau and Nagatsu (2020), and Jiménez-Buedo and Miller (2010) study the relationship between internal and external validity of experimental studies and the difficulties in securing causal knowledge. Like Cartwright, these philosophers – while making important contributions in improving our understanding about the RCT method and largely its design – often focus on one issue at a time.[3]

No paper in the existing philosophical literature has yet assessed the central question about the overall validity and causal claims of RCTs, which cannot be evaluated when studying an individual issue in isolation or just the method's design. This is only possible by providing a broad overview of the range of issues and problems that influence the estimated causal results and that arise not only in designing but also implementing and analysing trials in practice – with each additional issue generally further reducing the level of overall validity. The paper provides such a broad overview by pulling together a larger range of important methodological issues and constraints to estimating causal effects in trials. Some of these issues have not yet been thoroughly discussed in the existing philosophical literature on trials (ibid.), including for example the assumption that background traits remain constant during trial implementation or the assumption that all preconditions needed for the treatment to work are fully met – as outlined here. This paper thereby, on one hand, assesses a wider set of issues influencing causal results in RCTs simultaneously than in existing philosophical papers that tend to focus on important single issues that does not allow for assessing the overall validity of trials. On the other hand, it outlines the range of issues using a set of real-world studies – the ten most cited RCT studies[4] – that is not a common approach in studying the RCT method in existing philosophical papers but enables a broader perspective needed to assess overall validity.

The paper's focus and main contribution to the existing literature is assessing the overall validity of the RCT method and its causal claims, and providing a broader, practice-based account of causal bias in trials. The hope is that, in assessing an RCT, bias is no longer thought of as an individual bias (for example only as selection bias or sampling bias) but generally always as an RCT's overall (aggregate) bias. Though more difficult to measure, it is at least as large as any individual bias and better reflects the actual degree of a trial's overall bias. Examples from highly influential RCT studies are used with the aim of supporting this broader epistemological topic. Such breadth (the need to focus on the range of issues that together affect overall trial validity) inevitably comes at the cost of less depth on any single issue discussed within the paper. The paper builds on the existing work discussing methodological biases that can face RCTs by taking a new epistemological perspective to discuss and assess how these issues, together, can affect the causal claims and overall validity of trials (Krauss 2018; cf. Andrew et al. 1994; Black 1996; Sackett et al. 1996; Moher et al. 1998; Rennie 2001; Chan and Altman 2005; Dwan et al. 2008; Moher et al. 2010; Goldacre 2016; Richards et al. 2019).

This question of the overall validity of the causal claims of RCTs is important and at times a life-or-death matter in philosophy of science and medicine. RCTs, when the set of issues and problems outlined in the paper are taken together, generally produce a degree of bias in their estimated causal results that can influence their overall validity. The argument is that biases such as adherence bias, sampling bias and lack of blinding bias commonly affect, to some degree, the estimated causal results of trials, and as the ten most cited trials face these and other biases, most trials are bound to have some degree of bias and constrained overall validity. For the type of questions that fit the RCT design, the method is however often the best we have to inform our decisions and public policy. Yet given the complexities involved in studying human subjects, we should not assume that the RCT method always produces valid causal results. Such results cannot be viewed separately from the combined set of assumptions and biases underlying a given trial.

After this introduction, a background on trials is provided (Section 1). The common account of biased causal inference is then presented, and a broader, practice-based account of causal bias and validity is outlined (Sections 2 and 3). The long and complex path to causal claims in trials is then outlined – namely, the large set of issues and constraints are discussed that emerge when designing, implementing and analysing RCTs and that, together, affect their estimated causal results (Sections 4 and 5). Only when we assess the range of issues and problems facing trials, together, are we able to subsequently draw inferences about the *overall validity* of RCTs and their causal claims – which has not yet been done in the philosophical literature (Section 6). The paper then concludes outlining the implications on practice and on improving trials for the large communities of researchers using this method and for the small community of philosophers studying this method (Section 7).

1. A Brief Background on Trials

Across the biomedical, clinical, behavioural and social sciences causal relationships have (since the second half of the twentieth century) been increasingly estimated using the method of randomised controlled trials.[5] To answer the question about the causal effect of a treatment involves being able to compare what occurred with some treatment with what would have occurred without that treatment. This is not possible for an individual as we cannot assess a causal effect for the same person with and without the treatment at the same point in time. Researchers create a comparison or control group that would on average experience similar outcomes as those who are treated if they would not have been treated.

To this end, a randomised controlled trial randomly allocates people in the study sample into either a treatment group (who receive some medical treatment, technology, policy intervention or the like) or a control group (who are given a placebo, the common treatment at present, both or nothing at all). The aim is that the treatment would be the *only* causally relevant difference between the groups while other factors would be similar. Then the evaluator, blindfolded, statistically assesses the effectiveness of the treatment by comparing the average outcome in the experimental group to the average outcome in the control group. This allows assessing whether the treatment influenced the given outcome for at least one or more people (or units) in the experimental group compared to the

control group. Simply put, the difference between the mean outcomes in the treatment and control groups is seen as the estimated causal effect of the treatment.

The unique power of RCTs, of creating good counterfactuals and of producing reliable and valid causal results is commonly viewed to be found in randomisation (Ward and Johnson 2008; Worrall 2007, 2010; Papineau 1994; Harrison 2011). It is found in the random probability of being selected to receive a treatment or not – which can be thought of as the result of flipping a coin or rolling a dice. The division between randomised and non-randomised studies is viewed, for many, as the defining trait for scientific rigour and for deducing valid causal claims.[6] A number of the methodological and epistemological issues discussed throughout the paper are directly or indirectly related to randomisation. In the next section, we first clarify what is meant by causal bias and validity in trials, and outline the need for a broader account of causal bias than currently available in the literature.

2. From the Common Account of Causal Bias (as Bias Arising *Between* Trial Groups) Towards a Practice-based Account of Causal Bias (as Arising *Between*, *Within* and *Across* Trial Groups)

To identify the different aspects of a trial's design, implementation and analysis that contribute to bias, we need to understand how biases arise in causal inference and prediction. The leading account of biased causal inference in the RCT literature is grounded in the probabilistic theory of causality (Suppes 1970). That account of the causal foundations of RCTs describes bias in the following way: as RCTs 'allow causal claims about the population in the study to be deduced from probability differences between the treatment and control groups' (Cartwright 2007, 15; cf. Cartwright 1989; Heckman 2001; Holland and Rubin 1988), bias in causal inference is viewed as arising as a failure to ensure probabilistic independence of the treatment (cf. Cartwright 2010). That is, causal inference is not biased and 'an RCT is ideal iff all factors that can produce or eliminate a probabilistic dependence between C [the cause] and E [the effect] are the same in both wings except for C' – according to Cartwright (2010, 64). This is the common view of internal validity. It refers to the validity of a causal claim, and a causal claim is the average causal effect of a treatment on the individuals in the trial sample. That account provides important insights into the nature of biased causal inference and lays the theoretical foundation for understanding how probabilistic imbalance between trial groups can affect a trial's causal results, as outlined in Sections 4–6.

The common account of causal bias provides a general definition of bias, and does not aim to reflect the nature of different types of biases (Cartwright 2007, 2010). A broader, more detailed account of causal bias is provided here that builds on the common account to specify how the different sources of biases can arise. It helps to better reflect the actual practice of RCT design, implementation and analysis and is needed to help practitioners and policymakers know where to look for, identify and reduce diverse types of biases commonly affecting causal outcomes (Sections 4–6). The more idealised, common account operates at a higher level of abstraction than the more empirically-driven, practice-based account provided here that specifies the particular sources of biases and that is of direct use for practitioners and policymakers.

Sources of biases influence our estimated causal outcomes that are not made explicit in the common account of causal bias in terms of imbalances *between* trial groups (Cartwright 2007, 2010) but – as outlined in the practice-based account – arise *within* one of the trial groups or *across* the entire sample itself. For trials in which for example a share of participants *within only one* of the trial groups do not fully adhere to treatment dosages (or placebo dosages) and thus take different amounts, we have a source of biased causal inference (and thus prediction) not made explicit in that common account in terms of imbalances *between* trial groups. Causal bias thus arises when all individuals in the treatment group are not under the influence of C, so that for example a share of participants are not taking the full dosage of the given treatment for the full duration of the trial. This is often the case in practice (as outlined in Section 4). That is, bias arises in trials in which participants in the placebo group take the full dosage while a share of participants *only within* the treatment group do not adhere to taking the full dosage and differ in some way from those who do take the full dosage (or vice versa with a share of participants only within the placebo group). Such adherence bias is not covered by the common account of causal bias that only aims at covering imbalances in confounding factors – at ensuring that they 'are the *same in both wings* except for C' (Cartwright 2010, 64, emphasis added); but adherence bias instead arises during implementation *in only one wing*. So, we can refine and expand the common account of causal bias: causal inference is not biased and 'an RCT is ideal iff all factors that can produce or eliminate a probabilistic dependence between C [the cause] and E [the effect] are the same in both wings except for C' (Cartwright 2010, 64) and, *in addition*, the treatment group is homogeneous with respect to C (the treatment) and the placebo group is homogenous with respect to P (the placebo).[7]

For trials in which researchers or biopharmaceutical companies select the data points for the baseline and endline and thus choose to calculate one causal estimate instead of another, we can have another source of biased causal inference (and thus prediction) not made explicit in the common account in terms of imbalances *between* trial groups. This is because the unique time points selected may reflect the average, highest, lowest, no or another causal effect. A trial's estimated causal effects are biased if for example the baseline and endline data points, say after the treatment of a two-week exercise programme, illustrates limited causal effect, but the causal estimate is in fact high after six weeks that would be reflected if the baseline and endline data points would have been selected for say a three month period that collected data at multiple midline data points in order to assess variation in the causal estimates over time. This is an instance in which the causal estimate is biased *across* the entire sample as the treatment's estimated causal effect cannot be properly assessed and would not be of much, if any, use even for the participants in the sample. Multiple time points need to thus generally be collected and reported, over a proper period of time, to reduce bias in the particular causal estimate and to understand the trajectory of the causal estimate over time. So, we can further refine and expand the common account of causal bias: causal inference is not biased and 'an RCT is ideal iff all factors that can produce or eliminate a probabilistic dependence between C [the cause] and E [the effect] are the same in both wings except for C' (Cartwright 2010, 64) and, *in addition*, C is tested at multiple time points (and not at one potentially skewed time point). That is, what are called 'internally valid' trial results would, when such biases to the estimated causal results are strong, *not always be useful even for understanding*

the estimated effects for the subjects within the sample itself (let alone for the target population) (Section 3). For trials in which a share of participants *within* the control group visit a medical practitioner outside of the trial for example to try and establish whether they are receiving the placebo or the actual treatment, such as a cholesterol-lowering drug, and begin taking an over-the-counter cholesterol-reducing drug to alleviate their conditions, then estimated causal results face a source of within-group bias not made explicit in the common account of causal bias. Such sources of within-group and across-group bias, including a range of others, arise commonly when conducting trials in practice – as discussed in Sections 4–6.

The common account of RCT inference and bias reflects how bias can arise due to imbalances between groups (Cartwright 2007, 2010, 64) but, in practice, biases come about through different means at various steps when conducting trials including, more specifically, arising also within a group, and across the overall sample itself that affect the estimated causal outcomes. Causal bias is defined here in a more specific and nuanced way relevant for practice:

> Causal inference in an RCT is not biased if all confounding factors are equally distributed *between* the treatment and control groups, if all subjects receive the identical treatment *within* the treatment group and the identical placebo *within* the control group, and if all study design, implementation and analysis features *across* the entire sample are properly selected and carried out to produce an undistorted estimated average effect of the treatment (that is of use in practice).

The common account of causal bias (Cartwright 2010, 2007) can appear, for practitioners, particularly focused on trial design, and especially randomisation, in ensuring balance between groups at the beginning of a trial – as applies to trials in fields like economics, psychology, agriculture etc. It is thus not as helpful in providing guidance on identifying the different types of biases we face during trial design, implementation and analysis in practice and identifying how to mitigate them. The broader, practice-based account of causal bias offered here aims to help improve trials by directing our attention to the range of biases at play. It aims to better reflect complex experimental practice, looking specifically at the different aspects in carrying out trials that affect the estimated causal results (Sections 4–6). The ultimate purpose of evaluating the effectiveness of trials is to improve the lives of people in the real world. This is why researchers and practitioners need a broad understanding of how the range of various biases facing trials can arise.

Overall, conceiving bias in terms of differences between a trial's groups can distract from the broader and more important question of whether the overall causal estimate and overall trial is biased (due to distortions arising between, within or across groups) and whether it is of use in practice. A consequence of the practice-based account of causal bias offered here is the need for a broader view of overall validity.

3. A Broader View of *Overall Validity* in RCTs

The methodological debate around the validity of experimental studies is often structured in terms of a dichotomy between validity of the estimated results within the experimental setting (internal validity) and validity of the estimated results holding or

generalising outside of the experimental setting (external validity) (Campbell 1957; Campbell and Stanley 1963; Cook and Campbell 1979; Heukelom 2009; Jiménez-Buedo and Miller 2010; Reiss 2019; Favereau and Nagatsu 2020). Campbell (1957, 297) coined the terms and laid the foundation for distinguishing and assessing internal and external validity. He first defined internal validity as whether the experiment made a difference in the experimental context, and external validity as whether the difference (effect) can be extrapolated to other contexts and populations (ibid.; cf. Campbell and Stanley 1963). When a causal relationship between two variables is experimentally identified, the results are then often generalised beyond the experimental context (ibid.). This internal-external distinction is common across the economic, social and psychological sciences, while the efficacy-effectiveness distinction is common across the medical sciences. This is framed in terms of experimental trials estimating results under highly controlled conditions or under more real world conditions. In this context, Cartwright (2007a, 220; 2010) argues that 'in almost all cases there will be a trade off between internal validity and external validity […] The usual complaint here is about the artificiality of the circumstances required to secure internal validity' – with artificial meaning ideal (or highly controlled) conditions. The distinction is thus commonly viewed as a trade-off. The better we attempt to isolate the intervention from confounding factors and ensure it is driving the estimated causal effects, the less likely the results are representative of the intervention's effects in the target population (beyond the trial) in which conditions are not controlled for and other intervening factors (causes) can also operate (ibid; Campbell 1957; Campbell and Stanley 1963). Yet the distinction is not always to be viewed as a trade-off (cf. Heukelom 2009; Jiménez-Buedo and Miller 2010; Reiss 2019).

In practice, biases arise in trials that often affect, *simultaneously*, aspects related to both causal inference and prediction (aspects internal and external to an experiment). These include biases in trial implementation (such as selection bias, lack of blinding bias, adherence bias) and in research design (such as small and insufficient sample sizes) etc. (Sections 4–6). In light of the practice-based account of causal bias, ultimately all biases affecting a trial's validity, internally, will constrain us in extrapolating results, externally. When trials face such biases to their causal estimates, then causal inference (that refers to internal validity) and thus inevitably also causal prediction (that refers to external validity) are affected. Problems for internal validity are thus inevitably also problems for external validity. Think of trials for example in which baseline and endline data points are poorly selected, or the sample is collected in one particular clinic with only a small share of people willing to participate who have a particular background trait in common. In general, when a sample faces such biases within or across groups then regardless of how precise results may be estimated, the trial study is itself not always of much value in practice, even for studying and understanding the estimated effects on the select participants within the particular sample. Achieving internal validity, in such cases, can mean little by itself. Ultimately we do not run trials for their own statistical sake – i.e. with the aim of just measuring a causal estimate – independent of trial design, implementation and analysis. Clinical trials are run in order to use the results for individuals in the real world. Labelling such biases (like adherence bias, selection bias, lack of blinding bias, sampling bias etc.) as problems just of internal validity or just of external validity may often miss

the point for why trials are run in real-world practice. Bias arising in the design, implementation or analysis of a trial means biased results.[8]

The debate on distinguishing between mechanisms and correlations provides further evidence that it is not always possible to maintain a clear distinction between internal and external validity of RCTs, especially in medicine. Clarke et al. (2014, 347), who provide an important critique of evidence hierarchies in medicine, argue that 'The statistical 'set up' of an RCT is such that it maximises internal validity […] However, there is no a priori reason why the results of an RCT should be straightforwardly applicable to another population' – i.e. have external validity. These authors' main approach and claim is the following: 'we have divided evidence into evidence of correlation, such as is obtained from RCTs, observational studies and so on, and evidence of mechanisms, which is often obtained from laboratory experiments. We have argued that to evaluate a causal claim in medicine, evidence of mechanisms should be considered alongside evidence of correlation' – but also in order to 'address the problem of exporting the results of an RCT' (ibid.). However, despite their main claim, most of the about 2 million RCTs in medicine – indexed in the Cochrane Library (2021) – test a drug treatment, in which the mechanism is already embedded and thus how the effect comes about can be explained. For most cases in medicine a clear dichotomy between internal and external validity is not feasible. This is because phase 0, I and II trials are themselves small-scale experiments conducted in laboratories and clinics to test how the chemical property or substance of a medical drug works (the mechanism) among a smaller group of people. Phase III and IV trials are in turn larger scale experiments to test the effectiveness of that given medical drug (the average treatment effect) and long-term benefits. The mechanism is thus to some extent generally known and already embedded in the RCT design within the drug treatment.[9] It is precisely for this reason that trialists can acquire research funding for medical trials by indicating the chemical compound of the drug they are to test – and expected results – that requires an understanding of the mechanism at play. This important fact about trials is overlooked or not well understood in much of the philosophical RCT literature about mechanisms (cf. Clarke et al. 2014).

Validity is thus defined here in a more holistic way:

> The overall validity of an RCT and its causal claim is maximised if the range of causal biases that arise between, within and across trial groups (and generally affect validity both internally and externally simultaneously) are reduced as far as possible, if the RCT provides evidence of a mechanism and statistical correlation, and if the overall sample itself is generated randomly (and not just participants within an unrepresentative sample).

Such an integrated and practice-based view of causal bias and overall validity in trials aims to help improve the conditions of people to the greatest extent possible.

4. Issues Affecting Causal Claims and Overall Validity of Trials that Commonly Arise: Adherence Bias, Sampling Bias and Lack of Blinding Bias

Various assumptions and biases underlying RCTs cannot always be avoided – statistical and non-statistical techniques cannot fully eliminate them. In the following, three such examples are provided – adherence bias, sampling bias and lack of blinding bias – that affect the estimated causal results of trials, and are directly reflected in the broader, practice-based account of causal bias presented here.

In conducting any trial in practice, participants generally take the prescribed treatment for different lengths of time and different dosages that leads to adherence bias within one of the trial groups and affects the estimated causal results and overall validity. The length of follow-up for example was two or three times longer for some participants within half of the ten most cited RCTs discussed here – though only the average causal effect was estimated despite the different amounts of treatment received among different participants (DCC 1993; SSSSG 1994; Turner 1998; Knowler et al. 2002; Rossouw et al. 2002). Moreover, some share of participants generally do not take the intended dosage of the treatment – for example 27% did not take the intended dosage in the trial by Hurwitz et al. (2004) and 28% did not take at least 80% of the dosage in the trial by Knowler et al. (2002). This influences the estimated average causal results. Other statistical biases that influence our causal outcomes can also arise while implementing trials due to missing data for participants, participants switching between treatment and control groups and the like. In general, the more frequent such issues take place in each trial, the more problematic the causal and epistemic claims become. Some of these issues are difficult to fully detect while others become part of the statistical deviations. An example is that in the trial on estrogen by Rossouw et al. (2002), 42% of treated participants later discontinued the study drug, the vital status for 4% of participants was unknown (that is, data was missing) and 3% of participants passed away. Another difficulty in interpreting this trial's reported causal outcomes is that 11% of participants in the placebo arm switched to the active treatment arm. The individual choice made by participants to switch groups, once they become aware of which group they are in, must also be taken into consideration, especially participants' direct wellbeing, despite the fact that it can give rise to statistical bias. Beyond idealisations of the RCT method, such difficult to avoid methodological issues affect the top ten cited trial studies that do not provide details on the extent to which these issues bias their causal results and influence their overall validity.

*In conducting any trial in practice, some share of recruited people refuse to participate that leads to sampling bias through an overrepresentation of certain study participants and affects the estimated causal results and overall validity (*Banerjee and Duflo 2009*;* Kannisto et al. 2017*;* Heckman 2020*)*. Seven of the ten most cited trials do not report the share of people recruited who refuse to take part in the trial but, when reported for the remaining three trials, the share is high. In the trial by Shepherd et al. (1995), only 51% of those recruited for the trial showed up to the initial screening, of whom only 4% were then randomised into the trial. In the trial by SSSSG (1994), 8% of recruited people did not consent to take part. And in the trial by Rossouw et al. (2002), only 5% of all women initially screened consented to participate in the trial (and indicated not having had a hysterectomy), after which 88% of those who consented were then randomised into the study. The ideal view of trials is thereby that sampling bias can be eliminated but, in practice, only those participate who have time, expect to benefit, view minimal risk in taking the treatment and possibly have greater need for it (cf. Kannisto et al. 2017). An RCT does not capture these and other psychological factors that influence trials' estimated causal outcomes. That is, as people refuse to participate and thus the trial sample becomes smaller it is likely not 'average people' being lost but rather those who likely differ strongly (ibid.). Intention-to-treat analysis cannot fix such issues – or issues related to missing data etc. – that can lead to a

degree of causal bias *across the entire sample*. Average estimated causal effects in trials are thus likely generally biased upwards.

In conducting any trial in practice, some degree of partial blinding or unblinding of the various trial persons generally arises that leads to lack of blinding bias and affects the estimated causal results and overall validity. Blinding is needed throughout the trial because knowing which group (treatment or control) that participants are assigned to often biases decisions, intentionally or unintentionally, given different expectations and behaviour among practitioners and those treated (Teira 2013; Howick 2017). The ideal view of trials is however that randomisation, while helping to blind trial participants, can be sufficient to achieve blinded trials. Among the ten most cited trials, some were not double-blinded (Slamon et al. 2001; Van den Berghe et al. 2001; Hurwitz et al. 2004) whereas others were but later unblinded participants to allow for instance for management of adverse treatment effects (DCC 1993; Knowler et al. 2002; Rossouw et al. 2002). In the real world, in some cases it is not possible to carry out a blinded trial – with an example being the trial by Van den Berghe et al. (2001) in which participants' blood glucose levels must be continually monitored to be able to adjust insulin doses. In other cases, trial participants can unblind themselves – with an example being the trial by SSSSG (1994) in which some participants checked their cholesterol levels outside the trial and then discontinued the placebo to begin taking actual cholesterol-lowering drugs themselves. Moreover, these highly cited RCT studies did not provide details about how such issues of blinding and unblinding influence their reported causal and epistemic conclusions and overall validity. Yet to reduce such bias *within one of the trial groups* and increase the level of precision in the estimated causal results, trials need to be fully blinded (cf. Howick 2017) – though the top ten cited trials commonly only use (if at all) basic double-blinding. For relevant trials, full blinding implies that no one would know before, during or after the trial which participants were in the treatment and control group (Cartwright 2010), with no one meaning not only experiment designers and participants but also data collectors, practitioners, data evaluators or anyone else. RCT studies generally do not however report whether *all* these key people were blinded, as is the case in all top ten cited trials, which can increase uncertainty in the level of validity of their causal claims. Ultimately, when it comes down to estimating more precise causal results we need, for relevant studies, to get beyond basic Randomised Controlled Trials (RCTs) and move to fully Blinded and Randomised Controlled Trials (BRCTs).

In general, the causal claims of RCTs are thus commonly affected by such general issues of adherence bias, sampling bias and lack of blinding bias that can arise between, within and across trial groups. But causal claims are often influenced by a range of other methodological issues (Krauss 2018; cf. Andrew et al. 1994; Black 1996; Sackett et al. 1996; Moher et al. 1998; Rennie 2001; Chan and Altman 2005; Dwan et al. 2008; Moher et al. 2010; Goldacre 2016; Richards et al. 2019). These however need to be discussed, together, to assess overall validity and can also put into question the practical value of certain idealisations about trials.

5. Other Issues Affecting Causal Claims and Overall Validity of Trials

An assumption in trials is that the wide range of important background influencers needed for the treatment to work would exist simultaneously (the all-preconditions-are-fully-met

assumption of trials). We can however, in practice, only meaningfully estimate an intervention's causal effects (e.g. a medication, an education programme or the like) if recipients are sufficiently healthy and nourished for the intervention to function, if recipients take a sufficient amount of the prescribed intervention, if practitioners are qualified in carrying it out effectively, if the capacity of public institutions is adequate for its overall implementation, among other factors. No estimated causal results are thus affected solely by the intervention but by many other background attributes and conditions that can give rise to bias *between*, *within* or *across trial groups*. A number of these influence a treatment's estimated causal effects both within and outside a trial setting. That these and other such demanding preconditions (concauses) would be entirely satisfied for all participants is a foundational assumption implicit in the epistemic practice of randomised experimentation. In the real world, we are however not able to make sure that such concauses are present and evenly distributed among trial groups, since they are at times either known but we cannot easily collect data on them or may be unknown (Papineau 1994; Cartwright 2007a, 2010). They can thus contribute to a further degree of uncertainty in the level of validity of a trial's reported causal outcomes. Such background influencers (covariates) furthermore vary across and within different contexts, and they change over time. Because the degree to which they are met varies, the average estimated causal results across different samples also vary. Causation in practice, and the logic of experimental reasoning, need to thus be viewed more broadly than the particular disease or problem and its treatment.

The achieving-good-randomisation assumption is a fundamental assumption made in trials and in valuing their estimated causal outcomes and overall validity – which reflects the idealisation of randomisation and thus of being able to attain an even distribution of background traits that influence outcomes between trial groups. When trials are poorly randomised and face imbalances in background influencers between trial groups (Cartwright 2007, 2010; cf. Papineau 1994; Worrall 2007; Harrison 2011) it can lead to bias *between groups* and uncertainty about the validity of their estimated causal outcomes and epistemic conclusions. The common view that 'The RCT is neat because it allows us to learn causal conclusions without knowing what the possible confounding factors actually are' (Cartwright 2010, 64) does not fully reflect empirical practice as most leading journals require trial studies to provide baseline data with which one can generally observe some degree of imbalance in potential confounding factors between groups. An example among the top ten cited RCTs is the trial by Marler (1995) that suggests that the outcome is explained by the treatment (reflecting a 4 percentage point lower mortality rate for treated patients after three months of the stroke compared to placebo patients). When we disentangle the causal claims made in the trial it is not always possible to claim that this particular outcome is just caused by the treatment. This is because the trial's baseline data shows that background traits that shape the outcomes of stroke and mortality were not evenly distributed between trial groups: treated patients relative to placebo patients were on average 14% more likely to have taken aspirin therapy, 8% less likely to have been smoking, 3% less likely to have already had congestive heart failure, 3% more likely to be of white ethnicity compared to black, and 3% more likely to have already had (and survived) a stroke. Viewing these large differences in relation to the outcome of just a 4 percentage point difference in mortality, these alternative factors (causes) can also be explaining the trial outcomes. We cannot exclude the

possibility that the effect of the treatment may even have been negative and one or more of these alternative causes may be driving the outcomes. This poor randomisation and thus baseline distribution is however not discussed in the study. Another example of poor randomisation is the trial by Slamon et al. (2001) in which those in the treatment group (who received chemotherapy and the study treatment) were 10 percentage points more likely to have already had adjuvant chemotherapy prior to entering the trial relative to those in the control group (who only received chemotherapy). The epistemic practice of making causal claims in trials is complex because how people react to chemotherapy varies if they have received it before or not. So it may not be possible to claim that the treatment is exclusively affecting the trial's reported causal outcomes. For a given trial, a very balanced sample is not always easy to achieve because we use a finite sample with finite randomisations.[10] These most cited trials illustrate that randomisation does not guarantee an even allocation – meaning that the achieving-good-randomisation assumption may not hold for these RCTs to 'secure internal validity' and claim definitive causation about the treatment. In contrast to Cartwright's (2010, 63) view of the ideal RCT, random assignment cannot thus ensure, in real RCTs, 'that other possible reasons for dependencies and independencies between cause and effect under test will be distributed *identically* in the treatment and control wings' (emphasis added). Instead randomisation, when we do not critically assess such important imbalances, can also help shape a trial's estimated causal outcomes.[11] Observing that important differences in the actual frequencies exist, we cannot always exclude alternative explanations for the reported causal outcomes. Cartwright (2010), in contrast, theorises that evenly distributed probabilities may be sufficient independent of knowing the actual frequencies. Yet such a theory of randomisation does not reflect an important aspect of experimental practice because finite frequencies in trials are at times poorly distributed in practice (cf. Papineau 1994; Fuller 2018). In my view, the ultimate arbiter on the theoretical debate over whether probabilities or finite frequencies need to be evenly distributed is the real world we live in, i.e. empirical practice. And empirical practice (as seen above) illustrates that large and important imbalances in finite frequencies can arise and we cannot turn an eye to them as we know they also have an effect.[12]

In analysing outcomes, a further assumption in ideal trials is that a researcher's selected baseline and endline time points would properly reflect the average (or greatest possible) treatment outcome – yet, in practice, trials can face a unique-time-period-assessment bias. This is because a treatment's 'average' causal effect depends, for many trials, on when the time period is defined to gather the particular baseline and endline data. Making an epistemic claim about the estimated average causal outcome is a function of when the researcher assesses the treatment, whether every week or month, quarter or year etc. An example is that the trial by SSSSG (1994) estimates that the effect of the cholesterol treatment appeared to start after about a year and then it consequently decreased. Interventions tend to have varying effects at different points in time and may only work well in the shorter term but no longer, for example, once our body becomes accustomed to a certain medication. This brings an additional level of complexity to the epistemic practice of estimating causes that move over time. Overall, a treatment's evaluation will not be the same at different points in time and, for most trials, conducting assessments at multiple time points helps better understand the variations in the estimated causal results. Moreover, making an epistemic claim about the estimated

causal results in trials is a function of how the control group is designed. A placebo-only or conventional-treatment-only limitation of trials can thereby make it more difficult to interpret a trial's reported causal outcomes. Of the ten most cited trials, five evaluated the tested treatment only against the current treatment, four evaluated only against a placebo, and one evaluated against both to be able to compare relative causal outcomes using the two thresholds. One changes the causal question being addressed in each of these cases. We arrive at different answers to the same question of what is the impact of the treatment. In practice, such design features thus lead to different causal estimates *across the sample* and different epistemic conclusions – an insight not yet discussed in the philosophical literature, and not generally made explicit in trials.

The background-traits-remain-constant assumption is another assumption not yet discussed in the existing philosophical literature, though background traits change while a trial is conducted and they also influence estimated causal outcomes and overall validity so we need to always evaluate them not only at baseline but also at endline. When for example 5% of trial participants receiving the actual medication (compared to those receiving no or the conventional treatment) choose to simultaneously improve their physical fitness or diet in order to improve their conditions faster but we just gather baseline and not endline data on levels of physical fitness and diet, then we cannot claim that the trial's estimated causal results are just explained by the medication. And in general, as a trial period becomes longer background influences can often increasingly affect the estimated causal results given more potential confounders – for example related to changes in practitioners, clinic management, implementation of a national health policy etc. that take place after the trial begins. For such reasons, even a good baseline distribution does not eliminate potential bias *within one of the trial groups*. Randomisation cannot help in ensuring what happens post-randomisation. If researchers are not able to show that trial participants have the same background traits and clinics have the same characteristics at the endline that they had at the baseline, then they cannot know if the estimated causal outcome is only brought about by the intervention. Most RCTs face this concern (with none of the ten most cited RCTs having included such endline data) that contributes to a further degree of epistemic uncertainty in their level of bias and overall validity. The results of intention-to-treat analysis and per-protocol analysis can, for many trials, be thus thought of – to a certain extent – as both reflecting intention to treat. This is because such confounders after randomisation (and during trial implementation) are not being controlled for and trials are not estimating *total* treatment effects. This is an important insight into the actual epistemic practice of trials.

*In an ideal RCT all participants would moreover experience the same effect but, in practice, we face heterogeneity and outliers in data observations that can affect the estimated causal results in trials (*Deaton 2009*;* Ravallion 2009*;* Harrison 2011*).* Outliers and heterogeneity of treatment effects always exist because, for instance in medical trials, people are of different age, gender and physical health, experience different conditions, treatment needs and responses, develop different levels of resistance to the treatment etc. They are thus not a resolvable statistical or epistemic problem. They are a common consequence of studying dynamic biological, behavioural and social phenomena that involve complex processes. Such dynamic phenomena like diseases and economic policies are continually changing – their scope, their intensity, their duration etc. – and are better understood as (what I call) *evolving causes* rather than precisely measurable, static causes amenable to statistical

analysis. This can contribute to a further degree of uncertainty in the level of accuracy of a trial's causal estimates *across the entire sample*. Issues related to complexity, heterogeneity and evolving causes are however often not directly assessed as trials are designed specifically to estimate the average causal effect among the distribution.

Another assumption is that conditions of trial participants (within the controlled trial) would be the same or similar for those when the treatment may be later adopted in the general population (without experimental controls) which is widely discussed in the literature (e.g. Ward and Johnson 2008; Ravallion 2009; Cartwright 2010; Worrall 2010; Clarke et al. 2014; Marcellesi 2015; Reiss 2019; Favereau and Nagatsu 2020; Section 3). Influencing conditions include restrictive trial eligibility criteria, low participant consent, higher standards of trial facilities and trial practitioners etc. (Worrall 2007; Cartwright 2010). Such conditions in trials can at times give rise to bias *across the entire sample* by biasing upwards the estimated causal results. In practice, trials do not always evaluate in detail the central question of how their given causal results may be valid for individuals beyond the trial setting – as illustrated in most of these ten RCTs. Some however partly do. In the trial for example by Van den Berghe et al. (2001) conducted in one surgical care unit, its reported causal results are not applicable to individuals in medical care facilities or with illnesses different to those in the sample (which the study's authors recognise) but also to individuals with different demographic or clinical traits. Extrapolating results from one surgical care unit can thus raise important medical and epistemological concerns. Another example is that in the trial by Knowler et al. (2002), the authors seem to claim universal validity about their main reported causal outcome: 'To prevent one case of diabetes during a period of three years, 6.9 persons would have to participate in the lifestyle-intervention program, and 13.9 would have to receive metformin'. The authors then report that the trial's conclusions may however only be applicable to about 3% of the population in the US but also acknowledge that 'The validity of generalizing the results of previous prevention studies is uncertain. Interventions that work in some societies may not work in others, because social, economic, and cultural forces influence diet and exercise. This is a special concern in the United States, where there is great regional and ethnic diversity in lifestyle patterns' (Knowler et al. 2002). Overall, if trial studies do not elaborate in detail on their trial context and possible scope of their estimated causal outcomes beyond this context, then practitioners and policymakers must try and interpret the scope of the estimated causal outcomes themselves.

In general, in no two contexts are all conditions, needed for a treatment to have a positive causal effect, met to the same extent. It is unlikely that running the same trial (or scaling it up) in another context or in the same context at a different point in time would produce the same average causal effect. Because a trial's estimated causal results are relative to particular factors under particular conditions within a particular sample at a particular point in time, we can generally speak of a singular causal result. We can generally speak of a *one-time causal relationship*. Yet results derived from multiple studies increases the level of reliability and validity in experimental evidence by 'reducing' biases in individual trials.

Overall, while Cartwright (2010, 63–64) notes that 'in the design of real RCTs three features loom large' – blinding, random assignment and placebos – Sections 4 and 5 here illustrate that a range of other important features need to however be taken into

account simultaneously that bring assumptions and biases and influence trials' estimated causal results and overall validity (with those outlined above not being exhaustive).

6. Assessing the Overall Validity of Trials

It is, only when taking the large set of issues and problems arising in the design, implementation and analysis of trials together, then possible to assess the degree of overall validity of a trial and its causal claims. The question here is not about whether the set of assumptions must hold and biases must be eliminated for an RCT to establish causation. It is instead a question about the degree of validity of an RCT and its causal results being contingent on the degree to which we are able to reduce each bias and satisfy each assumption as far as possible. This central epistemological question about RCTs cannot be addressed studying just an individual issue in isolation or the method's design. It is because the level of validity of a trial can reduce with every additional bias and assumption that assessing overall validity has been constrained in papers focusing on a single issue. It has also been constrained given the common account of biased causal inference focused on imbalances between trial groups (Cartwright 2010, 2007). We return here to the broader, practice-based account of causal bias, outlined in Section 2:

> Causal inference in an RCT is not biased if all confounding factors are equally distributed *between* the treatment and control groups, if all subjects receive the identical treatment *within* the treatment group and the identical placebo *within* the control group, and if all study design, implementation and analysis features *across* the entire sample are properly selected and carried out to produce an undistorted estimated average effect of the treatment (that is of use in practice).

With this broader definition of causal bias (outlined in Sections 2 and 3) that better reflects the evidence of empirical practice of trials (outlined in Sections 4–6) it becomes evident that RCTs cannot generally reduce all biases which constrains their overall validity. What do the range of biases exactly affect? They impact the effect size, that is the strength of the relationship between the two variables that helps measure a study's practical significance. They also impact the statistical significance, that is the probability of the result (the difference measured) not arising due to chance for example at a significance level of 1 or 5 percent. The effect size in trials is influenced by adherence bias, lack of blinding bias, poor randomisation bias, a unique time period assessment bias, a change-in-background-traits bias etc. The statistical significance in trials is influenced by sampling bias, poor randomisation bias etc. (as outlined above). Some biases thus affect both.

Important statistical techniques, such as intention-to-treat analysis or stratified randomisation, have been devised to help mitigate the effects of a number of biases but cannot eliminate them. These techniques generally only reduce individual biases and for a number of biases discussed above there is no reliable way to assess how much they throw off results, and especially how much all combined biases affecting a given trial throw off results. Some of the studies discussed above face several (smaller) biases, for example related to adherence bias, sampling bias and lack of blinding bias, that generally add up to greater aggregate bias. But we face constraints in aggregating them in a statistically meaningful and precise way. It is nonetheless important to

assess and provide information in each trial, to the extent possible, on how much those different sources of bias, combined, can explain the estimated causal effect. In general, causal effects can face a greater degree of bias in trials with simultaneously smaller samples, smaller estimated effects, more unbalanced trial groups, a greater degree of partial blinding or unblinding etc. Improving our means to evaluate *aggregate bias* will be an important area for future research to improve trial methodology and the quality of research across fields using RCTs.

The response then to the larger question, how realistic is it to meet the set of assumptions and reduce the set of biases to establish precise causal results, is that we are generally dealing with some degree of bias in our causal claims – as illustrated assessing some of the most influential trials. Causal claims deduced in trials cannot be viewed separately from these assumptions and biases. The number of assumptions and biases implicit in an RCT's causal results and inferences can increase at each step: from selecting a sample, generating our variables, randomising and blinding, to implementing interventions, gathering data and dealing with unknown factors, interpreting our causal and epistemic conclusions, among other steps. Trial epistemology is highly complex. At least as many assumptions may thus exist in a given study as there are decisions and steps in carrying out the study. The causal claims we make based on our estimated results cannot be stronger than the weakest link used to arrive at those claims (see also Cartwright 2007) – they cannot be more valid than the strongest assumption we make or the strongest bias we face. Establishing which assumptions and biases are most and which least important in affecting the estimated causal results is not feasible in general terms. We can only assess their level of importance and the particular level of validity within a specific trial context and it is contingent on the degree to which the assumptions are met and the biases reduced. Biases such as adherence bias, sampling bias and lack of blinding bias commonly affect, to some degree, the estimated causal results of trials, and as the ten most cited trials face these and other biases, most trials are bound to have some degree of bias and constrained overall validity. This also applies more generally to other studies involving human beings that use statistics.

RCTs can moreover appear, in light of the common account of causal bias, at times as a one-step scientific method – focused on randomisation and creating balance between groups – to estimate causal relationships and ensure validity. In practice, to estimate its causal results an RCT rests on hundreds of decisions and steps that can lead to assumptions and biases before randomisation (e.g. sampling bias), during randomisation (e.g. poor randomisation bias) and after randomisation (e.g. lack-of-blinding or adherence bias) that arise between, within and across groups. Researchers thereby make decisions such as which particular factors are the most important to stratify for during randomisation. They choose how and in which location to select participants for their initial sample. They select and define which particular inclusion and exclusion criteria and outcome variables are appropriate. They choose the particular length of time between collected baseline and endline data points etc. Each of these decisions influence the particular causal results estimated in a trial. The single step of using a randomisation algorithm in the process of designing a trial cannot thus ensure causation, scientific objectivity or validity. It is, in practice, one step within the larger and highly complex process of designing, implementing and analysing the same trial.

Trials thus generally face some degree of biases that affect their causal results and overall validity – it is the character of the trade-off in order for studies to actually be carried out in the real world. A number of things do not always go according to plan (or design or theory) because designing, implementing and analysing studies and extrapolating results from them, in practice, is a long and intricate process involving many actors (study designers, participants, data collectors, implementing practitioners, statisticians etc.) making many unique decisions at many different steps over time. Any given study is prone to bias, limitations and assumptions, no matter how hard the researchers take measures to minimise them. Researchers conducting a study do not know what they do not yet know. And by the time a study is carried out and completed, flaws and biases become increasingly apparent over time. The question, for each study, is then whether the estimated causal outcomes with some degree of bias are sufficient for policy purposes and for practitioners? It generally is – though the response, for a particular trial, hinges on the level of overall validity and usefulness of its estimated outcomes in practice. If researchers and policymakers require having a gold standard in research or an evidence hierarchy it thus would not necessarily be tied to a particular method but, in my view, to the level of validity and reproducibility of research across methods and studies, and the usefulness of that research in practice. What is key here is conducting multiple studies that assess the same treatment and can help reduce the level of uncertainty in the estimated results and validity of each of the individual studies. Though, meta-analyses of trials do bring some of their own difficulties, such as the pooling together of biases from individual trials, likely overestimating causal effects of treatments and underestimating negative effects, since trials with negative results are less likely to get published.[13] Essential to improving the epistemic practice of trials is for researchers to provide much greater information, in their studies, on the assumptions they make and the methodological issues they face that influence their estimated causal effects. This is as important as any other information in allowing the reader to interpret the level of reliability of results, and derive credible conclusions.[14] If this critical information is not provided, readers cannot evaluate a study's overall validity.

In this sense, trials in fields like economics, psychology, political science and some areas of medicine are – in some regards and at times – not just a scientific tool but also a political tool. And one reason why some researchers in these fields view the experimental method of RCTs as the best way to improve understanding is because of the dynamic character of the phenomena studied and the weakness of theory in some of these fields. It is the RCT design that can provide greater rigour where theory may be lacking, especially in the behavioural and social sciences. In economics, RCTs are often used to justify public spending on a given intervention or policy. In medicine, they are often used as a policy tool to get new medical treatments passed through government regulatory bodies, and the like.

7. Conclusion

The philosophical literature on the RCT method has to date analysed important individual issues and biases (though largely in isolation) and it has adopted the common account of biased causal inference in which bias is described as arising due to probabilistic imbalances between trial groups (cf. Cartwright 2007, 2010). This paper has provided a

broader, practice-based account of causal bias in RCTs that incorporates the range of biases arising not only *between* groups but also *within* a group, and *across* the overall sample itself. Within this expanded framework, all biases are incorporated that arise in trial design, implementation and analysis that affect the estimated causal outcomes and overall validity. The paper has assessed a wider range of biases simultaneously, illustrating that an RCT generally faces a degree of *aggregate bias* that can constrain its causal claims and overall validity, but carrying out and comparing multiple studies helps 'mitigate' biases in individual studies. Overall validity increases as biases are reduced, and the robustness of our evidence increases as we conduct multiple studies.

RCTs are generally the best we have for the type of questions they can address. Randomised controlled trials are nonetheless not always as *random* as thought. For example, they are not always able to ensure an even distribution of all measurable, non-measurable and unknown influencers, and in many cases, the initial sample (of where and who to recruit) is not selected in a randomised way. They are also not always as *controlled* as thought. For example, they are not designed well to control for participants refusing to participate, not fully complying, dropping out etc. or, in the case of longer trials, for changes in background influencers during the trial. And they are not always *trials*. For example, some do not test a new intervention but rather only audit an existing (government) intervention.

Ultimately RCTs are not always assumption-free, bias-free and limitation-free. Yet we do not run RCTs because they guard against every bias and defect. We do them because they reduce biases and defects when experimenting. And it is because RCTs are vital in informing our policy decisions that it is so important for researchers to continually improve how trials are designed, implemented and analysed and improve their overall validity by continually identifying ways to reduce the degree of biases. Researchers can improve epistemic practice and patients' lives when they go through bias by bias and assumption by assumption, as outlined above, and try and minimise each bias and meet each assumption as far as possible when designing, implementing and analysing trials.

We need to also better combine RCTs with other methods – given that each method has its strengths – in order to gain a more holistic understanding of some phenomenon or treatment. Methods and research designs include single case studies[15] and laboratory methods (that are first steps needed to ground later experimentation), RCTs (that focus on the later stage of evaluation), observational studies (that help design and validate trials), consensus among groups of experts (that help resolve conflicting views), and at times 'historically controlled trials' (that can offer a historical perspective) especially in cases when RCTs cannot be conducted (Black 1996; Barbour 1999; Worrall 2007a, 2010; Ward and Johnson 2008; Clarke et al. 2014; Richards et al. 2019). Each method often provides different insights that RCTs are not able to and together they increase the robustness of evidence. For many questions we are interested in (from large-scale scientific topics to complex treatments, conditions, processes and institutions) RCT results cannot be more valid than results from other methods as we cannot generally apply them on such topics in practice. To address many topics not amenable to randomisation we require other methods such as observational studies but also to help in designing trials and interpreting and verifying their estimated causal outcomes and validity. For insights on the distribution of treatment effects among a population, why and how a treatment

can work (not only an average estimated causal effect), and under which conditions it can work, RCT results cannot generally be more valid (Deaton 2009; Ravallion 2009). This is because trials are not well designed for these purposes compared to other methods, such as at times observational studies. While historical and observational studies have at times been used, without RCTs, to identify important insights – from antibiotics and smoking inducing cancer, to surgical procedures and smallpox vaccination (Black 1996) – RCTs are often quicker and more efficient in providing causal knowledge, and together the different methods provide even stronger knowledge. The topics that fit an RCT design are largely those assessing, at the level of the individual, a single, simple, small-scale and quantifiable treatment with few known confounders. The reason why we cannot randomise most phenomena in science is generally because we do not have *a comparable enough counterfactual* for them – which draws the line of where and when trials can and cannot be applied. (Multiple) trials are – when implemented well, biases are reduced and assumptions are satisfied as far as possible – generally the best we have in better understanding the average estimated causal effect of a treatment and informing our decisions. No method is however all-around more valid than another method. The best all-around method does not exist and neither does a pre-established hierarchy of evidence (see also Clarke et al. 2014). We need to use RCTs together with other methods to provide insight into different aspects of a treatment or phenomenon that the different methods are not designed to do alone.

Finally, the paper has argued that to better explain and understand the complexities, inferences and biases in RCTs we require *a greater view from practice* and *a broader account of causal bias*. Only then can we gain a richer understanding about how trials produce some degree of bias – and thereby how to improve overall validity and how trials are carried out in practice.

Notes

1. For example, Andrew et al. 1994; Sackett et al. 1996; Djulbegovic et al. 2013.
2. For example, Seligman 1996; Duflo, Glennerster, and Kremer 2007; Banerjee 2007.
3. Holman (2017) also acknowledges some of the limitations of theoretical and idealised approaches to studying RCTs.
4. The examples provided throughout this paper are all found within the top ten cited RCT studies worldwide in any scientific journal. Each of these ten trials has been cited by at least 6,500 or more articles as of 2016 based on the Scopus database. They include world-leading trials on the topics of breast cancer (Slamon et al. 2001), colorectal cancer (Hurwitz et al. 2004), stroke (Marler 1995), postmenopause (Rossouw et al. 2002), insulin therapy (Van den Berghe et al. 2001), two separate trials on cholesterol (Shepherd et al. 1995; SSSSG 1994) and three separate trials on diabetes (Turner 1998; DCC 1993; Knowler et al. 2002). The issues discussed here, while these RCTs are a sample of highly influential trials and fall within medicine, biology and neurology, generally apply to any RCT across the medical, behavioural and social sciences and beyond; though there are differences in the design features of trials between fields like medicine, psychology and economics.
5. The RCT method yields what is often called an interventionist or manipulationist form of causation (Woodward 2003; see also Russo and Williamson 2011).
6. For a discussion on the meaning of causal claims in biomedical contexts, see Russo and Williamson (2011).
7. Cartwright refers here to balance 'in both wings', though balance is needed between all wings as trials in practice at times employ multiple treatment groups (to test different

treatments or dosages against each other) and multiple control groups (to test the treatment against both a placebo and the common treatment at present, for example).

8. The distinction of internal-external validity can also at times be problematic for another reason. There are many biases that often go beyond both, including industry sponsorship bias, reporting bias, reference bias (citing only studies favouring a given outcome), publication bias etc.
9. The majority of RCTs in medicine do not thus have what Clarke et al. (2014, 343) call the 'realistic chance of stumbling across coincidental correlations'.
10. In addition, computerised randomisation algorithms make the assumption that numbers can actually be selected entirely at random – but such algorithms, as Fallis (2000) argues, in fact tend to use a deterministic sequence in which initial values shape later values.
11. Researchers conducting trials in one field are not always aware of different design features in other fields. For example, in trials within economics all participants are commonly randomized in the entire sample before a trial begins, while participants are generally randomized on a roll-in basis in medical trials which can lead to greater imbalances in background traits of participants.
12. See Teira (2010) for a discussion on frequentist versus Bayesian clinical trials.
13. For a discussion on meta-analyses, see Moher et al. (1998), Stegenga (2011) and Holman (2018).
14. Most of the ten most cited trials were furthermore published after the *Consolidated Standards of Reporting Trials* guidelines were adopted (Andrew et al. 1994) – though these guidelines need to be extended to include the broader range of issues and constraints facing trials (cf. Moher et al. 2010; Rennie 2001).
15. Ankeny 2014.

Acknowledgements

I am grateful for comments from Nancy Cartwright, John Worrall, Federica Russo, Carl Hoefer, Bennett Holman, Stephan Guettinger, Federica Malfatti, Saana Jukola, Corinna Peters, Camille Lassale, Johannes Findl, Adrià Segarra, and anonymous journal reviewers. I received funding from the Marie Curie programme of the European Commission and the Beatriu de Pinós programme of the Government of Catalonia.

References

Andrew, E., A. Anis, T. Chalmers, M. Cho, M. Clarke, D. Felson, P. Gøtzsche, et al. 1994. "A Proposal for Structured Reporting of Randomized Controlled Trials." *JAMA* 272 (24): 1926–1931.

Ankeny, R. 2014. "The Overlooked Role of Cases in Casual Attribution in Medicine." *Philosophy of Science* 81 (5): 999–1011.

Banerjee, A. 2007. *Making Aid Work*. Cambridge: MIT Press.

Banerjee, A., and Esther Duflo. 2009. "The Experimental Approach to Development Economics." *Annual Review of Economics* 1 (1): 151–178.

Barbour, Rosaline. 1999. "The Case for Combining Qualitative and Quantitative Approaches in Health Services Research." *Journal of Health Services Research & Policy* 4 (1): 39–43.

Black, N. 1996. "Why we Need Observational Studies to Evaluate the Effectiveness of Health Care." *BMJ* 312 (7040): 1215–1218.

Campbell, D. 1957. "Factors Relevant to the Validity of Experiments in Social Settings." *Psychological Bulletin* 54: 297–312.

Campbell, D., and J. Stanley. 1963. *Experimental and Quasi-Experimental Designs for Research*. Chicago: Rand McNally and Company.

Cartwright, N. 1989. *Nature's Capacities and Their Measurement*. Oxford: Clarendon Press.

Cartwright, N. 2007a. *Hunting Causes and Using Them*. Cambridge: Cambridge University Press.

Cartwright, N. 2007. "Are RCTs the Gold Standard?" *Biosocieties* 2 (1): 11–20.
Cartwright, N. 2010. "What are Randomised Controlled Trials Good for?" *Philosophical Studies* 147: 59. https://doi.org/10.1007/s11098-009-9450-2.
Chan, A., and D. Altman. 2005. "Epidemiology and Reporting of Randomised Trials Published in PubMed Journals." *Lancet* 365: 1159–1162.
Clarke, B., D. Gillies, P. Illari, F. Russo, and J. Williamson. 2014. "Mechanisms and the Evidence Hierarchy." *Topoi* 33: 339–360.
Cochrane Library. 2021. *Cochrane Library*. London, https://www.cochranelibrary.com/central.
Cook, T., and D. Campbell. 1979. *Quasi-Experimentation: Design & Analysis Issues for Field Settings*. Boston: Houghton Mifflin Company.
Deaton, A. 2009. *Instruments of Development: Randomization in the Tropics, and the Search for the Elusive Keys to Economic Development*. NBER Working Papers 14690, National Bureau of Economic Research, Inc.
Diabetes Control and Complications Trial Research Group (DCC). 1993. "The Effect of Intensive Treatment of Diabetes on the Development and Progression of Long-Term Complications in Insulin-Dependent Diabetes Mellitus." *New England Journal of Medicine* 329 (14): 977–986.
Djulbegovic, B., A. Kumar, P. Glasziou, B. Miladinovic, and I. Chalmers. 2013. "Medical Research: Trial Unpredictability Yields Predictable Therapy Gains." *Nature* 500: 395–396.
Duflo, E., R. Glennerster, and M. Kremer. 2007. "Using Randomization in Development Economics Research: a Toolkit." In *Handbook of Development Economics*. Vol. 4, 3895–3962. Elsevier.
Dwan, K., D. G. Altman, J. A. Arnaiz, J. Bloom, A. W. Chan, E. Cronin, E. Decullier, et al. 2008. "Systematic Review of the Empirical Evidence of Study Publication Bias and Outcome Reporting Bias." *PLoS One* 3 (8): e3081.
Fallis, D. 2000. "The Reliability of Randomized Algorithms." *The British Journal for the Philosophy of Science* 51 (2): 255–271.
Favereau, Judith, and Michiru Nagatsu. 2020. "Holding Back from Theory: Limits and Methodological Alternatives of Randomized Field Experiments in Development Economics." *Journal of Economic Methodology* 27 (3): 191–211.
Fuller, J. 2018. "The Confounding Question of Confounding Causes in Randomized Trials." *The British Journal for the Philosophy of Science* 70 (3): 901–926.
Goldacre, B. 2016. "Make Journals Report Clinical Trials Properly." *Nature* 530 (7588): 7.
Gorard, Stephen, and Chris Taylor. 2004. *Combining Methods in Educational and Social Research*. Columbus, OH; Open University Press: McGraw-Hill.
Harrison, G. 2011. "Randomization and Its Discontents." *Journal of African Economies* 20 (4): 626–652.
Heckman, J. 2001. *Econometrics, counterfactuals and causal models*. Keynote Address, International Statistical Institute, Seoul, Korea.
Heckman, J. 2020. Randomization and Social Policy Evaluation Revisited. *Institute of Labor Economics*, IZA Discussion Papers, No. 12882, Bonn.
Heukelom, F. 2009. Origin and Interpretation of Internal and External Validity in Economics. *Nijmegen Center for Economics*; NiCE Working Paper 09-111.
Holland, P., and D. Rubin. 1988. "Causal Inference in Retrospective Studies." *Evaluation Review* 12: 203–231.
Holman, B. 2017. "Philosophers on Drugs." *Synthese* 196: 4363–4390.
Holman, B. 2018. "In Defense of Meta-Analysis." *Synthese* 196: 3189–3211.
Holman, B., and J. Bruner. 2017. "Experimentation by Industrial Selection." *Philosophy of Science* 84 (5): 1008–1019.
Howick, J. 2011. "Exposing the Vanities - and a Qualified Defense - of Mechanistic Reasoning in Health Care Decision Making." *Philosophy of Science* 78 (5): 926–940.
Howick, J. 2017. "The Relativity of 'Placebos': Defending a Modified Version of Grünbaum's Definition." *Synthese* 194: 1363–1396.

Hurwitz, H., L. Fehrenbacher, W. Novotny, T. Cartwright, J. Hainsworth, W. Heim, J. Berlin, et al. 2004. "Bevacizumab Plus Irinotecan, Fluorouracil, and Leucovorin for Metastatic Colorectal Cancer." *New England Journal of Medicine* 350: 2335–2342.

Jiménez-Buedo, María, and Luis Miller. 2010. "Why a Trade-off? The Relationship Between the External and Internal Validity of Experiments." *Theoria. Revista de Teoría, Historia y Fundamentos de la Ciencia* 25 (3): 301–321.

Kane, R. 2006. *Understanding Health Care Outcomes Research*. Burlington: Jones & Bartlett.

Kannisto, K. A., J. Korhonen, C. E. Adams, M. H. Koivunen, T. Vahlberg, and M. A. Välimäki. 2017. "Factors Associated With Dropout During Recruitment and Follow-Up Periods of a MHealth-Based Randomized Controlled Trial for Mobile.Net to Encourage Treatment Adherence for People With Serious Mental Health Problems." *Journal of Medical Internet Research* 19 (2): e46.

Kelly, A., R. Lesh, and J. Baek. 2014. *Handbook of Design Research Methods in Education: Innovations in Science, Technology, Engineering, and Mathematics Learning and Teaching*. New York: Routledge.

Knowler, W., Barrett-Connor E., Fowler S., Hamman R., Lachin J., Walker E., Nathan D. 2002. "Reduction in the Incidence of Type 2 Diabetes with Lifestyle Intervention or Metformin." *New England Journal of Medicine* 346 (6): 393–403.

Krauss, Alexander. 2018. "Why all Randomised Controlled Trials Produce Biased Results." *Annals of Medicine* 50: 312–322.

Marcellesi, A. 2015. "External Validity: Is There Still a Problem?" *Philosophy of Science* 82 (5): 1308–1317.

Marler, J. 1995. "Tissue Plasminogen Activator for Acute Ischemic Stroke." *New England Journal of Medicine* 333: 1581–1588.

Moher, D., S. Hopewell, K. F. Schulz, V. Montori, P. C. Gøtzsche, P. J. Devereaux, D. Elbourne, M. Egger, and D. G. Altman. 2010. "CONSORT 2010 Explanation and Elaboration: Updated Guidelines for Reporting Parallel Group Randomised Trials." *BMJ* 340: c869.

Moher, D., A. Jones, D. J. Cook, A. R. Jadad, M. Moher, P. Tugwell, and T. P. Klassen. 1998. "Does Quality of Reports of Randomised Trials Affect Estimates of Intervention Efficacy Reported in Meta-Analyses?" *Lancet* 352: 609–613.

Papineau, D. 1994. "The Virtues of Randomization." *The British Journal for the Philosophy of Science* 45: 437–450.

Ravallion, Martin. 2009. "Evaluation in the Practice of Development." *The World Bank Research Observer* 24 (1): 29–53.

Reiss, Julian. 2019. "Against External Validity." *Synthese* 196: 3103–3121.

Rennie, D. 2001. "CONSORT Revised – Improving the Reporting of Randomized Trials." *JAMA* 285: 2006–2007.

Richards, D. A., P. Bazeley, G. Borglin, P. Craig, R. Emsley, J. Frost, J. Hill, et al. 2019. " Integrating Quantitative and Qualitative Data and Findings When Undertaking Randomised Controlled Trials." *BMJ Open* 9 (11): e032081.

Rossouw, J. E., G. L. Anderson, R. L. Prentice, A. Z. LaCroix, C. Kooperberg, M. L. Stefanick, R. D. Jackson, et al. 2002. "Risks and Benefits of Estrogen Plus Progestin in Healthy Postmenopausal Women: Principal Results from the Women's Health Initiative Randomized Controlled Trial." *JAMA* 288:321–333.

Russo, F., and J. Williamson. 2011. "Epistemic Causality and Evidence-Based Medicine." *History and Philosophy of the Life Sciences* 33 (4): 563–581.

Sackett, D. L., W. M. Rosenberg, J. M. Gray, R. B. Haynes, and W. S. Richardson. 1996. "Evidence-based Medicine: What it is and What it Isn't." *British Medical Journal* 312: 71–72.

Scandinavian Simvastatin Survival Study Group (SSSSG). 1994. "Randomised Trial of Cholesterol Lowering in 4444 Patients with Coronary Heart Disease: the Scandinavian Simvastatin Survival Study." *Lancet* 344: 1383–1389.

Seligman, M. 1996. "Science as an Ally of Practice." *American Psychology* 51 (10): 1072–1079.

Shepherd, J., S. M. Cobbe, I. Ford, C. G. Isles, A. R. Lorimer, P. W. Macfarlane, J. H. McKillop, and C. J. Packard. 1995. "Prevention of Coronary Heart Disease with Pravastatin in men with Hypercholesterolemia." *New England Journal of Medicine* 333: 1301–1308.

Slamon, D. J., B. Leyland-Jones, S. Shak, H. Fuchs, V. Paton, A. Bajamonde, T. Fleming, et al. 2001. "Use of Chemotherapy Plus a Monoclonal Antibody Against her2 for Metastatic Breast Cancer That Overexpresses HER2." *New England Journal of Medicine* 344: 783–792.

Stegenga, J. 2011. "Is Meta-Analysis the Platinum Standard of Evidence?" *Studies in History and Philosophy of Biological and Biomedical Sciences* 42 (4): 497–507.

Suppes, P. 1970. *A Probabilistic Theory of Causality*. Amsterdam: North-Holland.

Teira, David. 2010. "Frequentist Versus Bayesian Clinical Trials." In *Philosophy of Medicine [Handbook of Philosophy of Science, vol. 16]*, edited by Fred Gifford, 255–297. Amsterdam: Elsevier.

Teira, D. 2013. "Blinding and the Non-Interference Assumption in Medical and Social Trials." *Philosophy of the Social Sciences* 43 (3): 358–372.

Turner, R. 1998. "Intensive Blood-Glucose Control with Sulphonylureas or Insulin Compared with Conventional Treatment and Risk of Complications in Patients with Type 2 Diabetes." *Lancet* 352: 837–853.

Van den Berghe, G., P. Wouters, F. Weekers, C. Verwaest, F. Bruyninckx, M. Schetz, D. Vlasselaers, P. Ferdinande, P. Lauwers, and R. Bouillon. 2001. "Intensive Insulin Therapy in Critically ill Patients." *New England Journal of Medicine* 345: 1359–1367.

Ward, A., and P. Johnson. 2008. "Addressing Confounding Errors When Using non-Experimental, Observational Data to Make Causal Claims." *Synthese* 163 (3): 419–432.

Waters, E., B. Swinburn, J. Seidell, and R. Uauy. 2010. *Preventing Childhood Obesity: Evidence Policy and Practice*. New Jersey: Wiley-Blackwell; BMJ books.

Woodward, J. 2003. *Making Things Happen*. Oxford: Oxford University Press.

Worrall, J. 2007a. "Evidence in Medicine and Evidence-Based Medicine." *Philosophy Compass* 2 (6): 981–1022.

Worrall, J. 2007. "Why There's no Cause to Randomize." *The British Journal for the Philosophy of Science* 58 (3): 451–488.

Worrall, J. 2010. "Evidence: Philosophy of Science Meets Medicine." *Journal of Evaluation in Clinical Practice* 16 (2): 356–362.