ANNALS OF MEDICINE, 2018
https://doi.org/10.1080/07853890.2018.1453233

Taylor & Francis
Taylor & Francis Group

ORIGINAL ARTICLE

OPEN ACCESS Check for updates

Why all randomised controlled trials produce biased results

Alexander Krauss

London School of Economics; University College London, London, UK

ABSTRACT

Background: Randomised controlled trials (RCTs) are commonly viewed as the best research method to inform public health and social policy. Usually they are thought of as providing the most rigorous evidence of a treatment's effectiveness without strong assumptions, biases and limitations.
Objective: This is the first study to examine that hypothesis by assessing the 10 most cited RCT studies worldwide.
Data sources: These 10 RCT studies with the highest number of citations in any journal (up to June 2016) were identified by searching Scopus (the largest database of peer-reviewed journals).
Results: This study shows that these world-leading RCTs that have influenced policy produce biased results by illustrating that participants' background traits that affect outcomes are often poorly distributed between trial groups, that the trials often neglect alternative factors contributing to their main reported outcome and, among many other issues, that the trials are often only partially blinded or unblinded. The study here also identifies a number of novel and important assumptions, biases and limitations not yet thoroughly discussed in existing studies that arise when designing, implementing and analysing trials.
Conclusions: Researchers and policymakers need to become better aware of the broader set of assumptions, biases and limitations in trials. Journals need to also begin requiring researchers to outline them in their studies. We need to furthermore better use RCTs together with other research methods.

KEY MESSAGES

- RCTs face a range of strong assumptions, biases and limitations that have not yet all been thoroughly discussed in the literature.
- This study assesses the 10 most cited RCTs worldwide and shows that trials inevitably produce bias.
- Trials involve complex processes – from randomising, blinding and controlling, to implementing treatments, monitoring participants etc. – that require many decisions and steps at different levels that bring their own assumptions and degree of bias to results.

ARTICLE HISTORY
Received 27 November 2017
Revised 10 January 2018
Accepted 13 March 2018

KEYWORDS
Randomised controlled trial; RCT; reproducibility crisis; replication crisis; bias; statistical bias; evidence-based medicine; evidence-based practice; reproducibility of results; clinical medicine; research design

Introduction

How well a given treatment may work can greatly influence our lives. But before we decide whether to take a treatment we generally want to know how effective it may be. Randomised controlled trials (RCTs) are commonly conducted by randomly distributing people into treatment and control groups to test if a treatment may be effective. Researchers in fields like medicine [1–4], psychology [5] and economics [6,7] often claim that this method is the only reliable means to properly inform medical, social and policy decisions; it is an ultimate benchmark against which to assess other methods; and it is exempt from strong theoretical assumptions, methodological biases and the influence of researchers (or as exempt as possible) which non-randomised methods are subject to.

This study assesses the hypothesis that randomised experiments estimate the effects of some treatment without strong assumptions, biases and limitations. In assessing this hypothesis, the 10 most cited RCT studies worldwide are analysed. These include highly influential randomised experiments on the topics of stroke [8], critically ill patients receiving insulin therapy [9], breast cancer and chemotherapy [10], estrogen and postmenopause [11], colorectal cancer [12], two trials on cholesterol and coronary heart disease [13,14] and three trials on diabetes [15–17]. While these trials are

CONTACT Alexander Krauss a.krauss@lse.ac.uk, alexander_krauss@hotmail.de London School of Economics; University College London, London, UK

related to the fields of general medicine, biology and neurology, the insights outlined here are as useful for researchers and practitioners using RCTs across any field including psychology, neuroscience, economics and, among others, agriculture.

This study shows that all of the 10 most cited RCTs assessed here suffer from at least several commonly known methodological issues that lead to biased results: poor allocation of their participants' background characteristics that influence outcomes across trial groups, issues related to partially blinding and unblinding, significant shares of participant refusal and participants switching between trial groups, among others. Some of these issues cannot be avoided in trials – and they affect their robustness and constrain their reported outcomes. This study thereby contributes to the literature on the methodological biases and limits of RCTs [1,18–25], and a number of meta-analyses of RCTs also indicate that trials at times face different biases, using common assessment criteria including randomisation, double-blinding, dropouts and withdrawals [20,21,26]. To help reduce biases, trial reporting guidelines [1,18] have been important but these need to be significantly improved.

A critical concern for trial quality is that only some trials report the common methodological problems. Even fewer explain how these problems affect their trial's results. And no existing trials report all such problems and explain how they influence trial outcomes. Exacerbating the situation, these are only some of the more commonly reported problems. This study's main contribution is outlining a larger set of important assumptions, biases and limitations facing RCTs that have not yet all been thoroughly discussed in trial studies.

Better understanding the limits of randomised experiments is very important for research, policy and practice. Even world-leading trials, while many help improve the conditions of those treated, all have at least some degree of bias in their estimated results and at times misguidedly claim to establish strong causal relationships. At the same time, some strongly biased trials are still used to inform practitioners and policymakers and can thus do harm for treated patients.

To be clear, the intention is not to isolate or criticise any particular RCTs. It is to stress that we should not trivialise and oversimplify the ability of the RCT method to provide robust conclusions about a treatment's average effect. Arriving at such conclusions is only possible if researchers go through each assumption and bias, one after the other (as outlined in this study), and make systematic efforts to try and meet these assumptions and reduce these biases as far as possible – while reporting those they are not able to.

Methods

This study selected trials using the single criterion of being one of the 10 most cited RCT studies. These 10 trials with the highest number of citations worldwide in any journal – up to June 2016 – were identified by searching Scopus (the largest database of peer-reviewed journals) for the terms "randomised controlled trial", "randomized controlled trial" and "RCT". These trials (each with 6500+ citations) were screened and each fulfilled the eligibility requirements of being randomised and controlled. For further information on the trial selection strategy and on the 10 most cited trials, see Appendix Figure A1 and Table 1.

This study, while applying and expanding common evaluation criteria for trials (such as randomisation, double-blinding, dropouts and withdrawals [20,21,26]), assesses RCTs using a broader range of assumptions, biases and limitations that emerge when carrying out trials. Terms I create for these assumptions, biases and limitations are placed *in italics*. In terms of the study's structure, the assumptions, biases and limitations are discussed together and in the order in which they arise in the design, then implementation, followed by analysis of RCTs.

Results and discussion

Assumptions, biases and limitations in designing RCTs

To begin, a constraint of RCTs not yet thoroughly discussed in existing studies is that randomisation is only possible for a small set of questions we are interested in – i.e. the *simple-treatment-at-the-individual-level limitation* of trials. Randomisation is largely infeasible for many complex scientific questions, e.g. on what drives overall good physical or mental health, high life expectancy, functioning public health institutions or, in general, what shapes any other intricate or large-scale phenomenon (from depression to social anxiety). Topics are generally not amenable to randomisation that are related to genetics, immunology, behaviour, mental states, human capacities, norms and practices. Not having a comparable counterfactual for such topics is often the reason for not being able to randomise. The method is constrained in studying treatments for rare diseases, one-off interventions (such as health system reforms) and interventions with lagged effects (such as treatments for long-term diseases). Trials are restricted in answering questions about how to achieve the desired outcomes within another context and policy setting: about what type of health practitioners are needed in which kind of clinics within what regulatory, administrative and institutional environment to deliver health services effective in providing the treatment. This method cannot, for such reasons, bring wholescale improvements in our general understanding of medicine. In cases where well-conducted RCTs are however most useful is in evaluating, for an anonymised sample, the average efficacy of a single, simple treatment assumed to have few known confounders – as published RCTs suggest. But they cannot generally be conducted in cases with multiple and complex treatments or outcomes simultaneously that often reflect the reality of medical situations (e.g. for understanding how to increase life expectancy or make public health institutions more effective). Researchers would, if they viewed RCTs as the only reliable research design, thus largely only focus on select questions related to simple treatments at the level of the individual that fit the quantifiable treatment–outcome schema (more to come on this later). They would let a particular method influence what type and range of questions we study and would neglect other important issues (e.g. increased life expectancy or improved public health institutions) that are studied using other methods (e.g. longitudinal observational studies or institutional analyses).

Another constraint facing RCTs is that a trial's initial sample, when the aim is to later scale up a treatment, would ideally need to be generated randomly and chosen representatively from the general population – but the 10 most cited RCTs at times use, when reported, a selective sample that can limit scaling up results and can lead to an *initial sample selection bias*. Some of these leading trials, as Table 1 indicates, do not provide information about how their initial sample was selected before randomisation [8,10] while others only state that "patient records" were used [13] or that they "recruited at 29 centers" [15]; but critical information is not provided such as the quality, diversity or location of such centres and the participating practitioners, how the centres were selected, the types of individuals they tend to treat and so forth. This means that we do not have details about the representativeness of the data used for these RCTs. Moreover, the trial on cholesterol by Shepherd et al. [14] was for example conducted in one district in the UK and the trial on insulin therapy by Van Den Berghe et al. [9] in one intensive care unit in Belgium – while both nonetheless aimed to later scale up the treatment broadly.

A foundational and strong assumption of RCTs (once the sample is chosen) is the *achieving-good-randomisation assumption*. Poor randomisation – and thus poor distribution of participants' background traits that affect outcomes between trial groups – puts into question the degree of robustness of the results from several of these 10 leading RCTs. The trial on strokes [8], which reports that mortality at 3 months after the onset of stroke was 17% in the treatment group and 21% in the placebo group, attributes this difference to the treatment. However, baseline data indicates that other factors that strongly affect the outcomes of stroke and mortality were not equally allocated: those receiving the main treatment (compared to those with the placebo) were 3% less likely to have had congestive heart failure, 8% less likely to have been smoking before the stroke, 14% more likely to have taken aspirin therapy, 3% more likely to be of white ethnicity relative to black, and 3% more likely to have had and survived a previous stroke. These factors can be driving the trial's main outcomes – in part or entirely. But the study does not explicitly discuss this very poor baseline allocation. In the breast cancer trial [10], 73% of treated participants (receiving chemotherapy plus the study treatment) had adjuvant chemotherapy before the trial compared to 63% of controlled participants (receiving chemotherapy alone). Because response to chemotherapy differs for those already exposed to it relative to those receiving it for the first time, it is difficult to claim that the study treatment was solely shaping the results. Likewise, the estimated main outcome of the colorectal cancer trial [12] – namely that those with treatment survived 4.5 months longer – cannot be viewed as a definitive result given that 4% more of those in the control group already had adjuvant chemotherapy. It is also unlikely that results in the diabetes trial by DCC [15] were not biased by the main intervention group having 5% less males, 2% more smokers and being 3% more likely to suffer from nerve damage. Some researchers may respond saying that "those may just be study design issues". But the point is that all of these 10 RCTs randomised their sample, showing that randomisation by itself does not ensure a balanced distribution – as we always have finite samples with finite randomisations. As long as there are important imbalances we cannot interpret the different outcomes between the treatment and control groups as simply reflecting the treatment's effectiveness. Researchers thus need to better reduce the degree of known imbalances – and thus biased results – by using larger samples, by selecting the most balanced distribution among multiple randomisation schedules and by stratified randomisation.

Another constraint that can arise in trials is when they do not collect baseline data for *all* relevant background influencers (but only some) that are known to alternatively influence outcomes – i.e. an *incomplete baseline data limitation*. These individual world-leading RCTs report for instance that heart disease reduced by taking the cholesterol-reducing drug called simvastatin [13] or the drug called pravastatin [14], that intensive diabetes therapy reduced complications of insulin-dependent diabetes mellitus [15], and that the duration that patients survive with colorectal cancer increased by taking the treatment called bevacizumab [12]. But these same trials do not collect baseline data – and thus assess – for differences between patients in levels of physical fitness, of exercise, of stress and other alternative factors that can also affect the primary outcome and bias results. The common claim, that "an advantage of RCTs is that nobody needs to know all the factors affecting the outcome as randomising should ensure it is due to the treatment", does not hold and we cannot evade an even balance of influencing factors.

When we observe, after randomising the sample, differences in the measurable influencing factors among the trial groups and if we for example re-randomise the same sample multiple times (before running the trial) until these factors are more evenly distributed, then we realise that trial outcomes are nonetheless the result of having only randomised once. We realise that trial outcomes would not be identical after each (re-)randomisation of the sample.

Moreover, some researchers argue that this method can minimise selection bias through blinded randomisation. Yet this can also be achieved by many other means of blinding. It is blinding, not randomising, that is crucial here [4]. For a trial to reduce selection bias and be completely blinded means that *nobody* – not just experimenters or patients but also data collectors, physicians, evaluators or anybody else – would know the group allocations. These 10 RCTs do not however provide explicit details on the blinding status of all these key trial persons throughout the trial.

Table 1 shows that some of these 10 trials did not double-blind [9,10,12] while others initially double-blinded but later partially unblinded [11,15,17] or only partially blinded for one arm of the trial [16] – which reflects in relevant cases (while often unavoidable) a *lack-of-blinding bias*. In the trial by Van Den Berghe et al. [9], for example, modifying insulin doses requires monitoring participants' glucose levels, making it impossible to run a blinded study. The estrogen trial [11] unblinded 40% of participants to allow for management of adverse effects. The diabetes trial by Knowler et al. [17] unblinded participants (though the share was not indicated) when their clinical results surpassed set thresholds and treatment needed to be changed. Some placebo patients in the trial by SSSSG [13] stopped the study drug to obtain actual cholesterol-lowering treatment which shows that treatment allocation was at times unblinded by participants themselves checking cholesterol levels outside the trial. Such issues related to blinding, although often unpreventable, need to be more explicitly discussed in studies and particularly the extent to which they bias results.

Beyond randomisation and blinding, a further constraint is that trials often consist of a few hundred individuals that are often too restrictive to produce robust results – i.e. the *small sample bias*. Among the top 10 RCTs, the two separate parts of the breast cancer trial [10] have sample sizes of 281 and 188 participants; and the two parts of the stroke trial [8] have sample sizes of 291 and 333 participants. Such small trials, together with at times strict inclusion and exclusion criteria and poor randomisation, often bring about important imbalances in background influencers and bias results (as shown earlier for these two studies) [21]. Small trials also face other large problems. An example is that the stroke trial [8] with 624 participants reports that at 3 months after the stroke, 54 treated patients died compared to 64 placebo patients. This main outcome is the same likelihood as getting 10 more heads than tails by flipping a coin 624 times. For a trial that say has 400 participants, when those treated are 3% more likely to achieve an outcome, this just means having the same probability of getting 206 heads in 400 random flips of a coin. Overall, to increase reliability in estimated results researchers ideally need large samples (if possible, thousands of observations across a broad range of different groups with different background traits) that estimate large effects across different studies. This would furthermore ideally be combined with more studies comparing different treatments against each other within a single trial – and testing (in relevant cases) multiple combined treatments in unison [e.g. comparing (i) increased exercise, (ii) improved nutrition, (iii) no smoking, (iv) a particular medication etc. in one trial with different treatments to assess relative benefits: (i), (i + ii), (i + ii + iii) and (i + ii + iii + iv)].

Another issue facing RCTs not yet discussed in existing studies is the *quantitative variable limitation*: that trials are only possible for those specific phenomena for which we can create strictly defined outcome variables that fit within our experimental model and make correlational or causal claims possible. The 10 most cited RCTs thus all use a rigid quantitative outcome variable. Some use the binary treatment variable (1 or 0) of whether participants died or not [9,12,13]. But this binary variable can neglect the multiple ways in which participants perceive the quality of their life while receiving treatment. In the colorectal cancer trial [12], for example, the primary outcome is an average longer survival of 4.5 months for those treated; but they were also 11% more likely to suffer grade 3 or 4 adverse events, 5% more likely to be hospitalised for such adverse events and 14% more likely to experience hypertension. These variables for adverse effects are nonetheless proxies and do not perfectly capture patients' quality of life or level of pain which are, by their very character, not directly amendable to quantitative analysis. Only using the variables captured in the trial, we do not have important information about whether participants who lived several months longer – but also suffered more intensely and longer – may have later preferred no treatment. Another example of the quantitative variable limitation is that the diabetes trial by Knowler et al. [17] sets the treatment as the goal of at least 150 min of physical activity per week. This treatment with a homogenous threshold nonetheless neglects factors that influence the effects of 150 min of exercise and thus the estimated outcomes – factors such as inevitable variation in participants' level of physical fitness before entering the trial and in their physiological needs for different levels of physical activity that depend on their specific age, gender, weight etc. This clear-cut quantitative variable (while often the character of the RCT method) thus does not reflect the heterogeneous needs of patients and decisions of practitioners. In fact, most medical phenomena (from depression, cancer and overall health, to medical norms and hospital capacity) are not naturally binary or amendable to randomisation and statistical analysis (and this issue also affects other statistical methods and its implications need to be discussed in studies).

Assumptions, biases and limitations in implementing RCTs

An assumption in implementing trials that has not yet been thoroughly discussed in existing studies is the *all-preconditions-are-fully-met assumption*: that a trial treatment can only work if a broad set of influencing factors (beyond the treatment) that can be difficult to measure and control would be simultaneously present. A treatment – whether chemotherapy or a cholesterol drug – can only work if patients are nourished and healthy enough for the treatment to be effective, if compliance is high enough in taking the proper dosage, if community clinics administering the treatment are not of low quality, if practitioners are trained and experienced in delivering it effectively, if institutional capacity of the health services to monitor and evaluate its implementation is sufficient, among many other issues. The underlying assumption is that all these and other such preconditions – causes – would be fully met for all participants. Ensuring that they are all present and balanced between trial groups, even if the sample is large, can be difficult as such factors are at times known but non-observable or are unknown. Variation in the extent to which such preconditions are met leads to variation (bias) in average treatment effects across different groups of people. To increase

Table 1. Research designs of the ten most cited RCTs worldwide

	Study reported							*Reported participants'*									
Trial	Initial sample selection	Eligibility criteria	Exclusion criteria	Refusal rate	Randomised stratification	Double-blinded	Even # of participants betw. treatment and control groups	Non-compliance rate (during implementation)	Drop-out rate	Reported multiple time points of collected data	Assessed back-ground traits at endline	Reported some adverse effects (not only positive)	Discussed alternative factors that affect main outcome	Reported degree of 'external validity' of study results	Reported research assumptions, biases and limitations	Sample size	Citations
Insulin-dependent diabetes [15]	No[i]	Yes	No	No	By intervention cohorts at each clinical centre	Partially[iii]	No	No	< 1%	Yes	No	Yes	No	Yes	No	1,441	16,279
Intensive blood-glucose control and type 2 diabetes [16]	Yes	Yes	Yes	No[ii]	By ideal bodyweight, and some patients by two kinds of treatment	Partially[iv]	No	No	4%	Yes	No	Yes	No	No	No	3,867	13,788
Estrogen and postmenopause [11]	Partially	Yes	Yes	95%	By clinical centre and age group	Partially[iii]	No	No	42%	Yes	No	Yes	No	Yes	Partially	16,608	10,792
Cholesterol and coronary heart disease [13]	No[i]	Yes	Yes	8%	By clinical centre and previous myocardial infarction	Yes	No	5% stopped taking drug	12%	Yes	No	Yes	No	No	No	4,444	9,659
Type 2 diabetes and lifestyle intervention [17]	Yes	Yes	Yes	No[ii]	By clinical centre	Partially[iii]	No	72% took ≥ 80% of dosage	8%	Yes	No	Yes	No	Yes	Partially	3,234	9,581
Colorectal cancer [12]	No[i]	Yes	Yes	No	By clinical centre, baseline treatment response status, location of disease and # of metastatic sites	No	No	73% took intended dosage	Partially (8% due to adverse effect)	Yes	No	Yes	No	No	No	813	7,025
Acute ischemic stroke [8]	No	Yes	Yes	No[ii]	By clinical centre and time between stroke and treatment	Yes	No	90-93% (±5) took intended dosage	No[ii]	Yes	No	Yes	No	No	No	291 *and* 333	6,839
Cholesterol and coronary heart disease [14]	Yes	Yes	Yes	≥49%[i]	By clinical centre and time of recruitment	Partially[v]	No	No[i]	30%	Yes	No	Yes	No	Partially	No	6,595	6,624
Insulin for ill patients [9]	Yes	Yes	Yes	No[ii]	By type of critical illness	No	No	No	No	n.a.[vi]	No	No	No	Yes	No[i]	1,548	6,582
Breast cancer and chemotherapy [10]	No	Yes	Yes	No	Insufficient information provided	No	No	92% took ≥ 80% of dosage	Partially (8% due to heart failure)	Yes	No	Yes	No	No	No	469	6,533

Source: Own illustration. Note: Number of citations reflects up to June 2016. [i]Study insufficiently reported information. [ii]Study did not explicitly report information. [iii]Study was initially double-blinded but later partially unblinded. [iv]Study only double-blinded one arm of the trial. [v]Study did not blind trial statistician. [vi]Study only reported a single time point as one surgery was conducted (not multiple). For further details on any given item in the table, see the respective section throughout the study.

the effectiveness of treatments and the usefulness of results, researchers need to give greater focus, when designing trials and when extrapolating from them, to this broader context.

In these 10 leading RCTs, some degree of statistical bias arises during implementation through issues related to people initially recruited who refused to participate, participants switching between trial groups, variations in actual dosage taken, missing data for participants and the like. Table 1 illustrates that for the few trials in which the share of people unwilling to participate after being recruited was reported it accounted at times for a large share of the eligible sample. Among all women screened for the estrogen trial [11], only 5% provided consent for the trial (and reported no hysterectomy). This implies a selection bias among those who have time, are willing, find it useful, view limited risk in participating and possibly have greater demand for treatment. Among this small share, 88% were then randomised into the trial. During implementation, 42% in the treatment group stopped taking the drug. Among all participants 4% had unknown vital status (missing data) and 3% died. As a sample gets smaller due to people refusing, people with missing data etc. "average participants" are likely not being lost but those who may differ strongly – which are issues that intention-to-treat analysis cannot necessarily address. A constraint in interpreting the estrogen trial's results is that 11% of placebo participants crossed over to the treatment arm. Decisions to switch between groups, once patients become familiar with the trial, need to also be understood in terms of their immediate health and lives – not just in terms of the statistical bias it brings to results.

One of the two cholesterol trials [14] reported that 51% recruited to participate appeared for the first screening, after which only 4% of the recruited sample was randomised into the study – and later about 30% of participants dropped out. In the other cholesterol trial [13], 8% of those eligible did not consent to participate, while 12% later stopped the drug due to adverse effects but also reluctance to continue. Non-compliance also arises in several of these RCTs. In one of the diabetes trials [17], the share of participants taking at least 80% of the prescribed dosage was 72% for those in the treatment group. In the colorectal cancer trial [12], 73% in the treatment group took the intended dose of one of the drugs. That significant shares of participants in these and other trials have different levels of treatment compliance (see Table 1) can lead to variation (bias) in estimating outcomes across participants (whether using intention-to-treat or per-protocol analysis). Also, several of these trials did not provide complete data on dropout rates (Table 1). Among them is the stroke trial [8] and for all participants with missing outcome data "the worst possible score was assigned". This assumption is not likely correct. Overall, what decisions researchers take to deal with participant refusal, switching between groups, missing data etc. raises difficult methodological issues and a further degree of bias in results that researchers need to openly discuss in trial studies.

Assumptions, biases and limitations in analysing RCTs

In evaluating results after trial implementation, RCTs face a *unique time period assessment bias* that has not yet been thoroughly discussed in existing studies: that a correlational or causal claim about the outcome is a function of when a researcher chooses to collect baseline and endline data points and thus assesses one average outcome instead of another. The trial by SSSSG [13] for example reports that the effect of the cholesterol-lowering drug seemed to begin, on average, after about a year and then subsequently reduced. Treatments generally have different levels of decreasing (or at times increasing) returns. Variation in estimated results is thus generally inevitable depending on when we decide to evaluate a treatment – every month, quarter, year or several years. No two assessment points are identical and we need to thus evaluate at multiple time points to improve our understanding of the evaluation trajectory and of lags over time (while this issue also affects other statistical methods).

In half of these 10 RCTs, the total length of follow-up was not always identical but at times two or three times longer for some participants – though these studies just reported the average results [11,13,15–17]. Different time lengths or different amounts of doses, however, bring about different effects between trial participants and can lead to biased results. In the trial on breast cancer and chemotherapy [10] for example, participants in the primary treatment group remained in the study between 1 and 127 weeks (on average 40 weeks) and the doses taken ranged between 1 and 98 (on average 36 doses).

Another strong assumption made in evaluating RCTs that has not yet been discussed is the *background-traits-remain-constant assumption* – but these change during the trial so we need to assess them not only at baseline but also at endline as they can alternatively influence outcomes and bias results. The longer the trial is the more important these influences often become. But they are also important for shorter trials: if those in the control group are given the common treatment or nothing at all and for example 3% of those in the treatment group decide to combine the tested drug treatment with other forms of treatment such as additional exercise or better nutrition to improve their conditions more rapidly but we only collect baseline and not endline data on levels of exercise and nutrition, then we do not know if the tested drug treatment alone is driving the outcomes. Unless we can ensure that participants at the endline have the identical background conditions and clinic traits that they had at the baseline, we cannot claim that "the outcome is just because of the treatment". This issue applies to all 10 RCTs as they do not include such endline data.

Another constraint is that trials are commonly designed to only evaluate average effects – i.e. the *average treatment effects limitation*. Though, average effects can at times be positive even when some or the majority are not influenced or even negatively influenced by the treatment but a minority still experience large effects.

Most of these top 10 RCTs, which have the objective to use the treatment in a broader population, do not fully assess how the results may apply to people outside the trial (Table 1) – i.e. the *extrapolation limitation*. A few however do partially report this information. The trial by Shepherd et al. [14] for example states that their results could be "applicable to typical middle-aged men with hypercholesterolemia". But it does not indicate if the results would only apply to typical men in the particular sub-population within the West of Scotland (where the trial was run) given the specific lifestyle, nutrition and other traits of people in this region and the capacity of participating clinics. For the trial by Van Den Berghe et al. [9], participants were selected for insulin therapy in one surgical intensive care unit. This implies that results cannot be applied to those in medical intensive care units or those with illnesses not present in the sample (which the authors acknowledge) but also to those with different demographic or clinical traits. One trial [11] explicitly reported not to use the tested treatment (estrogen) due to the health risks exceeding possible gains. The diabetes trial by Knowler et al. [17] provides most detail on the study's applicability compared to other top 10 trials, conceding 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 [for example] diet and exercise". The authors of this trial state that the results could apply to about 3% of the US population. In general, when researchers however do not explicitly discuss the potential scope of their results outside the trial context, practitioners do not exactly know whom they may apply to.

A *best results bias* can also exist in reporting treatment effects, with funders and journals at times less likely to accept negligible or negative results. Of these 10 trials, researchers at times indicate possible alternative explanations (beyond the treatment) for adverse treatment effects (e.g. in the colorectal cancer trial [12]). But these 10 trials do not explicitly discuss other measurable or non-measurable confounders, like the imbalanced background traits outlined above, that also shape the main (treatment) outcome (Table 1). Only one of these trials (the estrogen trial [11]) had a negative main treatment effect. The trial by Van Den Berghe et al. [9] did not discuss the adverse effects of the insulin therapy, but only reported an extensive list of its benefits.

Another constraint in evaluating trials is that funders can have some inherent interest in the published outcomes that can lead to a *funder bias*. This has been shown by a number of systematic reviews of trials [27,28]. Among the ten most cited RCTs, seven were financed by biopharmaceutical companies. The colorectal cancer trial [12] was funded and designed by the biotech company Genentech and it collected and analysed the data, while the researchers also received payments from the company for consulting, lectures and research. This was also the case for the breast cancer trial [10]. However, drug suppliers should not ideally, because of commercial interests, be independently involved in trial design, implementation and analysis – with one potential source of bias emerging through the selection of an inappropriate comparator to the tested treatment [27,28].

An associated constraint that arises in interpreting a trial's treatment effects is related to a *placebo-only or conventional-treatment-only limitation*. Four of the 10 trials compare the treatment under study only with a placebo [8,11,13,14] which can, in relevant cases, make it more difficult to inform policy as we do not know how the tested treatment directly compares with the current or conventional treatment. Five of the 10 trials compare the treatment only with conventional treatments [9,10,12,15,16] (and not additionally with a placebo) though a treatment's reported benefit can at times be attributed to the poor outcome in the conventional group. Only one of these trials [17] was designed for assessing the relative benefit of the tested and conventional treatments comparatively against a placebo.

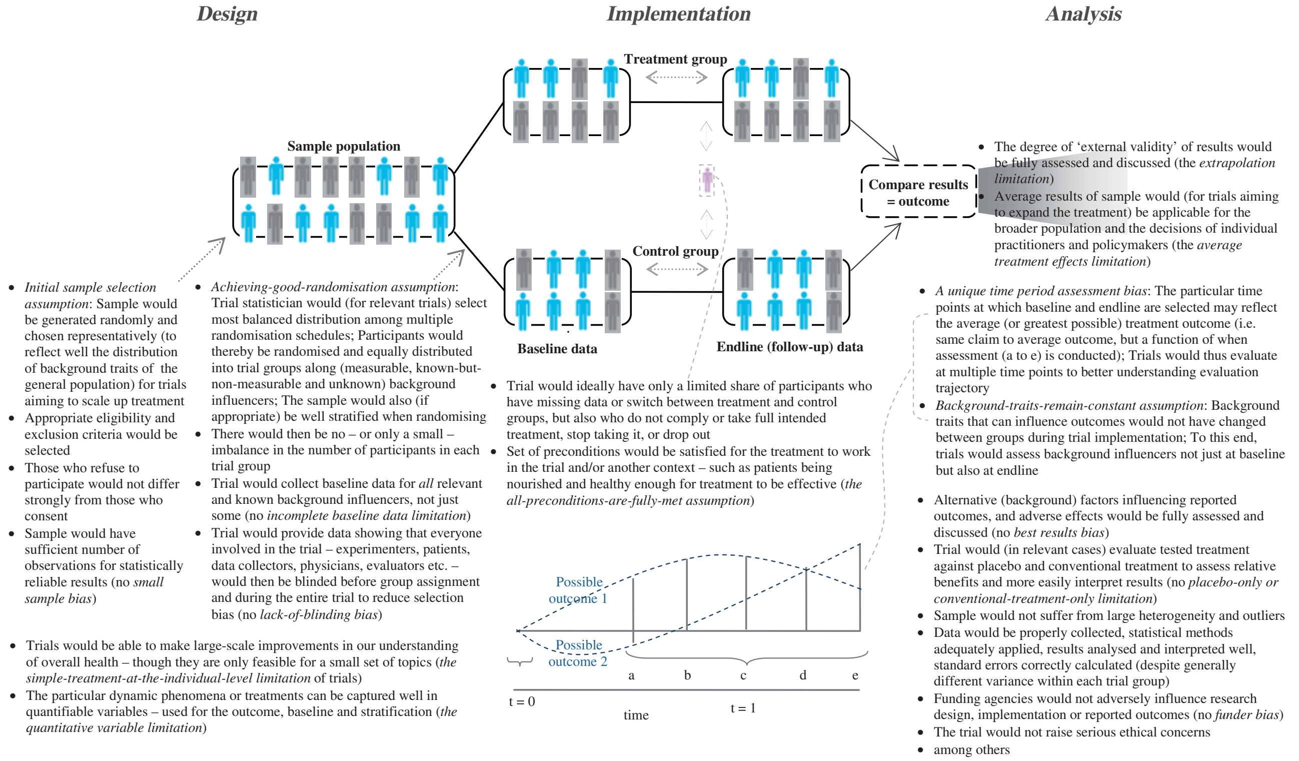
A number of other biases and constraints can also arise in conducting RCTs. These range from calculating standard errors (with the number of participants between trial groups being uneven in all 10 trials, as Table 1 illustrates), placebo effects [29], variations in the way sample sizes are determined, in the way different enumerators collect data for the same trial, and in the methods used to create the random allocation sequence, to differences in analysing, interpreting and reporting statistical data and results, changes in the design or methods after trials begin such as exclusion criteria, conducting subgroup analysis (and related ex post data-mining) [30], ethical constraints [31], budgetary limitations, and much more.

Combining the set of assumptions, biases and limitations facing RCTs

Pulling the range of assumptions and biases together that arise in designing, implementing and analysing trials (Figure 1), we can try to assess how reliable an RCT's outcomes are. This depends on the degree to which we may be able to meet each assumption and reduce each bias – which is also how researchers can improve trials.

Yet is it feasible to always meet this set of assumptions and minimise this set of biases? The answer does not seem positive when assessing these leading RCTs. The extent of assumptions and biases underlying a trial's results can increase at each stage: from how we choose our research question and objective, create our variables, select our sample, randomise, blind and control, to how we carry out treatments and monitor participants, collect our data and conduct our data analysis, interpret our results and do everything else before, in between and after these steps. Ultimately our results can be no more precise than such assumptions we make and biases we have. It is, in general terms, not possible to talk about which of them are more important. That can only be assessed in a given trial and depends on the extent to which each assumption is satisfied and each bias reduced.

We need to furthermore use RCTs together with other methods that also have benefits. When a trial suggests that a new treatment can be effective for some participants in the sample, subsequent observational studies for example can often be important to provide insight into: a treatment's broader range of side effects, the distribution of effects on those of different age, location and other traits and, among others, whether people in everyday practice with everyday service providers in everyday facilities would be able to attain comparable outcomes as the average trial participant. Single case studies and methods in



- *Initial sample selection assumption*: Sample would be generated randomly and chosen representatively (to reflect well the distribution of background traits of the general population) for trials aiming to scale up treatment
- Appropriate eligibility and exclusion criteria would be selected
- Those who refuse to participate would not differ strongly from those who consent
- Sample would have sufficient number of observations for statistically reliable results (no *small sample bias*)

- *Achieving-good-randomisation assumption*: Trial statistician would (for relevant trials) select most balanced distribution among multiple randomisation schedules; Participants would thereby be randomised and equally distributed into trial groups along (measurable, known-but-non-measurable and unknown) background influencers; The sample would also (if appropriate) be well stratified when randomising
- There would then be no – or only a small – imbalance in the number of participants in each trial group
- Trial would collect baseline data for *all* relevant and known background influencers, not just some (no *incomplete baseline data limitation*)
- Trial would provide data showing that everyone involved in the trial – experimenters, patients, data collectors, physicians, evaluators etc. – would then be blinded before group assignment and during the entire trial to reduce selection bias (no *lack-of-blinding bias*)

- Trials would be able to make large-scale improvements in our understanding of overall health – though they are only feasible for a small set of topics (*the simple-treatment-at-the-individual-level limitation* of trials)
- The particular dynamic phenomena or treatments can be captured well in quantifiable variables – used for the outcome, baseline and stratification (*the quantitative variable limitation*)

- Trial would ideally have only a limited share of participants who have missing data or switch between treatment and control groups, but also who do not comply or take full intended treatment, stop taking it, or drop out
- Set of preconditions would be satisfied for the treatment to work in the trial and/or another context – such as patients being nourished and healthy enough for treatment to be effective (*the all-preconditions-are-fully-met assumption*)

- The degree of 'external validity' of results would be fully assessed and discussed (the *extrapolation limitation*)
- Average results of sample would (for trials aiming to expand the treatment) be applicable for the broader population and the decisions of individual practitioners and policymakers (the *average treatment effects limitation*)

- *A unique time period assessment bias*: The particular time points at which baseline and endline are selected may reflect the average (or greatest possible) treatment outcome (i.e. same claim to average outcome, but a function of when assessment (a to e) is conducted); Trials would thus evaluate at multiple time points to better understanding evaluation trajectory
- *Background-traits-remain-constant assumption*: Background traits that can influence outcomes would not have changed between groups during trial implementation; To this end, trials would assess background influencers not just at baseline but also at endline

- Alternative (background) factors influencing reported outcomes, and adverse effects would be fully assessed and discussed (no *best results bias*)
- Trial would (in relevant cases) evaluate tested treatment against placebo and conventional treatment to assess relative benefits and more easily interpret results (no *placebo-only or conventional-treatment-only limitation*)
- Sample would not suffer from large heterogeneity and outliers
- Data would be properly collected, statistical methods adequately applied, results analysed and interpreted well, standard errors correctly calculated (despite generally different variance within each trial group)
- Funding agencies would not adversely influence research design, implementation or reported outcomes (no *funder bias*)
- The trial would not raise serious ethical concerns
- among others

Figure 1. Overview of assumptions, biases and limitations in RCTs (*i.e. improving trials involves reducing these biases and satisfying these assumptions as far as possible*). Source: Own illustration. Note: For further details on any assumption, bias or limitation, see the respective section throughout the study. This list is not exhaustive.

and outside of the laboratory are furthermore essential first steps that ground later experimentation and make later evaluation using RCTs possible. Moreover, to attain some of the medical community's most significant insights, historical and observational methods were used and RCTs were not later needed (and at times not possible), ranging from most surgical procedures, antibiotics and aspirin, to smallpox immunisation, anaesthesia, immobilising broken bones, smoking inducing cancer, among many other examples [23].

Conclusions

Randomised experiments require much more than just randomising an experiment to identify a treatment's effectiveness. They involve many decisions and complex steps that bring their own assumptions and degree of bias before, during and after randomisation. Seen through this lens, the reproducibility crisis can also be explained by the scientific process being a complex human process involving many actors making many decisions at many levels when designing, implementing and analysing studies, with some degree of bias inevitably arising during this process. And addressing one bias can at times mean introducing another bias (e.g. making a sample more heterogeneous can help improve how useful results are after the trial but can also reduce reliability in the trial's estimated results).

We then have to always make a judgement: are biased results in studies good enough to inform our decisions? Often they are – but that judgement generally depends on how useful the results are in practice and their level of robustness compared with other studies using the same method or, at times, other methods. Yet no single study should be the sole and authoritative source used to inform policy and our decisions. In general however, the impact of RCTs would be greater if researchers would systematically go through and aim to reduce each bias and satisfy each assumption as far as possible – as outlined in Figure 1. More broadly, what are the lessons for researchers to improve RCTs?

Journals must begin requiring that researchers include a standalone section with additional tables in their studies on the "*Research assumptions, biases and limitations*" they faced in carrying out the trial. Each trial should thereby have to include a separate table with the information listed in the CONSORT guidelines that have to be significantly expanded to also require not yet reported information on the share, traits as well as reasons of participants refusing to participate before randomisation, not taking full dosages, having missing data etc., on the blinding status of *all* key trial persons, on alternative (background) factors that can affect the main outcome and on the wider range of issues discussed throughout this study (Figure 1). It needs to also include a table with endline data (not just baseline data) of participants' background traits and clinic characteristics – and also more detailed information on the "*applicability of results*" including the broader range of background influencers of participants, step-by-step information on how the initial sample is exactly generated (not just eligibility criteria and clinic location) and whom the trial results may explicitly apply to. These 10 RCTs do not discuss all such essential information and the particular assumptions, biases and limitations (Table 1) – nor do they include all the information already in the CONSORT guidelines while most of these trials were published after the standardised international guidelines were agreed upon [1]. This study here thus highlights, on one hand, wider issues such as not fully understanding study reporting guidelines or not fully complying with the guidelines for minimally robust trials. It also raises the important question of why a number of high-profile studies that do not match up to minimal quality standards and have biased results continue to be highly cited. On the other hand, it illustrates that the CONSORT guidelines must be greatly extended to reflect this larger set of assumptions, biases and limitations. If journals begin requiring these additional tables and information (e.g. as an online supplementary appendix due to word limits), researchers would learn to better detect and reduce problems facing trials in design, implementation and evaluation – and thus help improve RCTs. Without this essential information in studies, readers are not able to assess well a trial's validity and conclusions. Some researchers may respond saying that they may already be familiar with a number of the biases outlined here. That however does not always seem to be the case as otherwise these influential RCTs would not all suffer, to such an extent, from some of these biases.

Researchers need to furthermore better combine methods as each can provide insight into different aspects of a treatment. These range from RCTs, observational studies and historically controlled trials, to rich single cases and consensus of experts. Some researchers may respond, "are RCTs not still more credible than these other methods even if they may have biases?" For most questions we are interested in, RCTs cannot be more credible because they cannot be applied (as outlined above). Other methods (such as observational studies) are needed for many questions not amendable to randomisation but also at times to help design trials, interpret and validate their results, provide further insight on the broader conditions under which treatments may work, among other reasons discussed earlier. Different methods are thus complements (not rivals) in improving understanding.

Finally, randomisation does not always even out everything well at the baseline and it cannot control for endline imbalances in background influencers. No researcher should thus just generate a single randomisation schedule and then use it to run an experiment. Instead researchers need to run a set of randomisation iterations before conducting a trial and select the one with the most balanced distribution of background influencers between trial groups, and then also control for changes in those background influencers during the trial by collecting endline data. Though if researchers hold onto the belief that flipping a coin brings us closer to scientific rigour and understanding than for example systematically ensuring participants are distributed well at baseline and endline, then scientific understanding will be undermined in the name of computer-based randomisation.

Acknowledgements

I am thankful for comments from Stephan Guettinger, Corinna Peters, Richard Baxter, Alina Velias, Nicolas Friederici, Severin Pinilla, Ravi Somani, Gunnar Wegner, Adrián Garlati and anonymous journal reviewers.

Disclosure statement

No potential conflict of interest was reported by the author.

Funding

This work was supported by the European Union under the Marie Sklodowska-Curie programme [grant # 745447].

References

[1] Andrew E, Anis A, Chalmers T, et al. A proposal for structured reporting of randomized controlled trials. JAMA. 1994;272:1926–1931.

[2] Sackett D, Rosenberg W, Gray J, et al. Evidence-based medicine: what it is and what it isn't. BMJ. 1996;312:71–72.

[3] Djulbegovic B, Kumar A, Glasziou P, et al. Medical research: trial unpredictability yields predictable therapy gains. Nature. 2013;500:395–396.

[4] Worrall J. Why there's no cause to randomize. London: London School of Economics; Nov 2004. (Technical report 24/4).

[5] Seligman M. Science as an ally of practice. Am Psychol. 1996;51:1072–1079.

[6] Duflo E, Glennerster R, Kremer M. Using randomization in development economics research: a toolkit. In: Handbook of development economics. Amsterdam: Elsevier; 2007.

[7] Banerjee A. Making aid work. Cambridge: MIT Press; 2007.

[8] Marler J. Tissue plasminogen activator for acute ischemic stroke. N Engl J Med. 1995;333: 1581–1588.

[9] Van Den Berghe G, Wouters P, Weekers F, et al. Intensive insulin therapy in critically ill patients. N Engl J Med. 2001;345:1359–1367.

[10] Slamon D, Leyland-Jones B, Shak S, et al. Use of chemotherapy plus a monoclonal antibody against her2 for metastatic breast cancer that overexpresses HER2. N Engl J Med. 2001;344:783–792.

[11] Rossouw J, Anderson G, Prentice R, et al. Risks and benefits of estrogen plus progestin in healthy postmenopausal women: principal results from the women's health initiative randomized controlled trial. JAMA. 2002;288:321–333.

[12] Hurwitz H, Fehrenbacher L, Novotny W, et al. Bevacizumab plus irinotecan, fluorouracil, and leucovorin for metastatic colorectal cancer. N Engl J Med. 2004;350:2335–2342.

[13] Scandinavian Simvastatin Survival Study Group (SSSSG). Randomised trial of cholesterol lowering in 4444 patients with coronary heart disease: the scandinavian simvastatin survival study. Lancet. 1994;344: 1383–1389.

[14] Shepherd J, Cobbe S, Ford I, et al. Prevention of coronary heart disease with pravastatin in men with hypercholesterolemia. N Engl J Med. 1995;333:1301–1308.

[15] Diabetes Control and Complications Trial Research Group (DCC). The effect of intensive treatment of diabetes on the development and progression of long-term complications in insulin-dependent diabetes mellitus. N Engl J Med. 1993;329:977–986.

[16] Turner R. Intensive blood-glucose control with sulphonylureas or insulin compared with conventional treatment and risk of complications in patients with type 2 diabetes. Lancet. 1998;352:837–853.

[17] Knowler W, Barrett-Connor E, Fowler S, et al. Reduction in the incidence of type 2 diabetes with lifestyle intervention or metformin. N Engl J Med. 2002;346:393–403.

[18] Moher D, Hopewell S, Schulz K, et al. CONSORT 2010 Explanation and Elaboration: updated guidelines for reporting parallel group randomised trials. BMJ. 2010;340:c869.

[19] Rennie D. CONSORT revised – improving the reporting of randomized trials. JAMA. 2001;285:2006–2007.

[20] Moher D, Pham B, Jones A, et al. Does quality of reports of randomised trials affect estimates of intervention efficacy reported in meta-analyses? Lancet. 1998;352:609–613.

[21] Chan A, Altman D. Epidemiology and reporting of randomised trials published in PubMed journals. Lancet. 2005;365:1159–1162.

[22] Vandenbroucke J. Observational research, randomised trials, and two views of medical science. PLoS Med. 2008;5:e67.

[23] Black N. Why we need observational studies to evaluate the effectiveness of health care. BMJ. 1996;312: 1215–1218.

[24] Dwan K, Altman D, Arnaiz J, et al. Systematic review of the empirical evidence of study publication bias and outcome reporting bias. PLoS One. 2008; 3:e3081.

[25] Goldacre B. Make journals report clinical trials properly. Nature. 2016;530:7.

[26] Lawler P, Filion K, Eisenberg M. Efficacy of exercise-based cardiac rehabilitation post-myocardial infarction: a systematic review and meta-analysis of randomized controlled trials. Am Heart J. 2011;162:571–584.e2.

[27] Bekelman J, Li Y, Gross C. Scope and impact of financial conflicts of interest in biomedical research: a systematic review. JAMA. 2003;289: 454–465.

[28] Lexchin J, Bero L, Djulbegovic B, et al. Pharmaceutical industry sponsorship and research outcome and quality: systematic review. BMJ. 2003;326:1167–1170.

[29] Allison M. Reinventing clinical trials. Nat Biotechnol. 2012;30:41–49.

[30] Yusuf S, Wittes J. Interpreting geographic variations in results of randomized, controlled trials. N Engl J Med. 2016;375:2263–2271.

[31] Pfeffer M, McMurray J. Lessons in uncertainty and humility – clinical trials involving hypertension. N Engl J Med. 2016;375:1756–1766.

Appendix

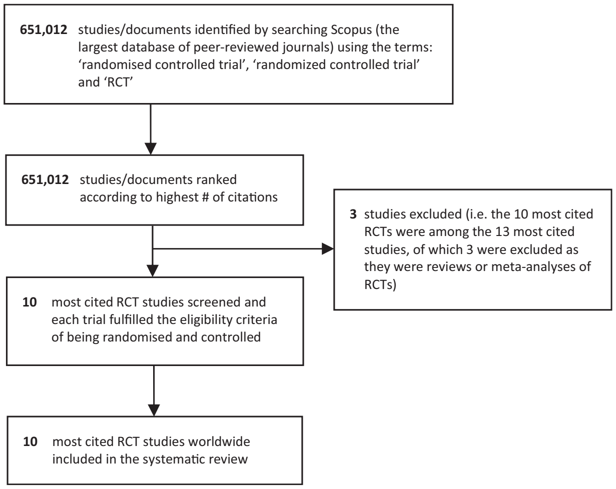

651,012 studies/documents identified by searching Scopus (the largest database of peer-reviewed journals) using the terms: 'randomised controlled trial', 'randomized controlled trial' and 'RCT'
651,012 studies/documents ranked according to highest # of citations
3 studies excluded (i.e. the 10 most cited RCTs were among the 13 most cited studies, of which 3 were excluded as they were reviews or meta-analyses of RCTs)
10 most cited RCT studies screened and each trial fulfilled the eligibility criteria of being randomised and controlled
10 most cited RCT studies worldwide included in the systematic review

Figure A1. PRISMA flowchart – selection of studies for the review. Source: Own illustration. Note: RCT studies selected based on number of citations up to June 2016.