Institute for Fiscal Studies

Randomisation bias in the medical literature: a review

IFS Working Paper W16/24

Barbara Sianesi

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL

‘Randomisation bias’ in the medical literature: A review

November 2015

Barbara Sianesi

Institute for Fiscal Studies

Abstract: Randomised controlled or clinical trials (RCTs) are generally viewed as the most reliable method to draw causal inference as to the effects of a treatment, as they should guarantee that the individuals being compared differ only in terms of their exposure to the treatment of interest. This ‘gold standard’ result however hinges on the requirement that the randomisation device determines the random allocation of individuals to the treatment without affecting any other element of the causal model. This ‘no randomisation bias’ assumption is generally untestable but if violated would undermine the causal inference emerging from an RCT, both in terms of its internal validity and in terms of its relevance for policy purposes. This paper offers a concise review of how the medical literature identifies and deals with such issues.

JEL codes: C18, C21, C90, 19

Keywords: Clinical trials, social experiments, design of experiments, randomisation bias, sample selection, causal inference, treatment effects, external validity, generalizability

Acknowledgments: This review has arisen whilst revising Sianesi (2014). I thank a referee for suggesting I explore the medical literature on this topic.
Financial support from the European Research Council (ERC) under grant agreement No. 269440-WSCWTBDS and from the ESRC Centre for the Microeconomic Analysis of Public Policy at IFS under grant reference ES/M010147/1 is gratefully acknowledged.

Address for correspondence:
Institute for Fiscal Studies, 7 Ridgmount Street, WC1E 7AE London, UK.
E-mail: barbara_s@ifs.org.uk.

1. Introduction

Randomised clinical trials (RCTs) are the recognised gold standard underpinning evidence-based medicine, as they ensure that groups of subjects being compared differ only in terms of their exposure to the treatment or intervention under investigation. Randomisation is a device introduced purely for evaluation purposes; the identifying and generally untestable assumption of the experimental design is thus that randomisation determines the random allocation of individuals but affects no other element of the causal model, an assumption which in econometrics has been referred to as 'no-randomisation bias' (Heckman, 1992). One of the most powerful critiques to the use of randomised trials is indeed the possibility that individuals might react to the randomisation itself, thereby rendering the causal inference from the trial either invalid or inappropriate for policy purposes.

In the medical field, scholars have been aware of "issues related to the process of randomisation that may affect the validity of conclusions drawn from the results of RCTs" (Britton *et al.*, 1998, p.iii), with some authors (e.g. Bartlett *et al*., 2005, and Bornhöft *et al*., 2006) clearly highlighting the tension between features that safeguard internal validity (such as the randomisation itself, placebo and blinding) but at the same time worsen external validity or indeed generate or exacerbate what we call randomisation bias.

In this paper we first outline the theoretical framework set out in Sianesi (2014) to systematically consider randomisation bias, in other words, how randomisation *per se* may affect potential outcomes and participation choices and thus give rise to misleading causal inference. We subsequently map the discussion in the clinical literature into this framework, concisely reviewing how the medical literature has identified and attempted to deal with these issues. We conclude with a brief comparison to the econometrics literature.

2. A theoretical framework for randomisation bias

Following Sianesi (2014)[1], consider a treatment (e.g. a healthcare intervention) which can be administered either in routine mode or within a randomised trial. The parameter of interest is the average treatment effect on the treated (*ATT*). Randomisation bias is present if implementing the intervention alongside a randomised trial gives rise to an average treatment effect on the treated which is *different* from the average effect on the treated which would have arisen had the treatment been administered in routine mode. Box 1 formalises our

[1] I am indebted to a referee for suggestions on how to set up a theoretical framework to think about randomisation bias.

definition of randomisation bias.

Box 1: Formal characterization of randomisation bias

Randomisation bias is present if:

$$ATT(1) \equiv E[Y_{1i}(1) - Y_{0i}(1) \mid D_i(1)=1] \neq E[Y_{1i}(0) - Y_{0i}(0) \mid D_i(0)=1] \equiv ATT(0)$$

where

- $ATT(RCT)$ is the average treatment effect on the treated when the intervention is administered in routine mode (RCT=0) or within a trial (RCT=1);
- $Y_{0i}(RCT)$ and $Y_{1i}(RCT)$ are the potential outcomes of individual i under no-treatment and under treatment, where such potential outcomes are allowed to depend on whether the treatment is administered in routine mode (RCT=0) or via a trial (RCT=1); and
- $D_i(RCT)$ is an indicator denoting participation in the intervention by individual i, where such a decision can potentially also be affected by the evaluation mode of the intervention.

There are thus two main channels through which randomization bias may operate: randomisation *per se* may affect impacts via affecting potential outcomes Y_0 and/or Y_1, and it may affect participation choices D.

The most serious threat to an RCT is the possibility that the no-treatment outcome has been affected by randomisation. In economics, we typically focus on the likelihood of *control group contamination* (or *substitution*), whereby control group members engage in a different type or intensity of alternative programs from what they would have done in the absence of the experiment. Such a situation would irremediably compromise the internal validity of the trial, as the control group would no longer provide the correct no-treatment counterfactual for the program group.

Treatment outcomes Y_1 may be affected by random assignment e.g. when randomisation disrupts the normal operation of the program. This type of randomisation bias would typically affect the external validity of the estimate, as we would obtain unbiased estimates of the effect of the disrupted intervention.

Finally, randomisation bias would be present if

(a) randomisation has a causal average effect on trial participation choices D among the eligibles, i.e. $E[D(1)–D(0)] \neq 0$, so that the proportion of participants differs under randomization and in routine mode, i.e. $Prob(D(1)=1) \neq Prob(D(0)=1)$, *and*

(b) such differences between the populations participating under the two scenarios translate into a difference in the corresponding ATT's. For randomisation bias to occur, a necessary condition is thus that treatment impacts vary across patients, i.e.

there is treatment effect heterogeneity.

According to potential participation behaviour, the population of all eligibles can be partitioned into four groups (similarly to the *LATE* framework of Imbens and Angrist, 1994): the *defiers* are those who decide not to participate because of randomisation, the *compliers* are those who are induced to participate by randomization, the *always-takers* would participate in any case and the *never-takers* would not participate under either scenario.

When all that is available is a randomised trial, we can only observe the participants under randomisation (the $D(1)=1$ group made up of the always-takers and the compliers) and the effect for them, while both the group of treated under normal operation (the $D(0)=1$ group, encompassing the always-takers and the defiers) and the treatment effect for this group ($ATT(0)$) are in general unobserved as they entail some counterfactual components. Given that the relevant parameters to make a decision about whether or not to implement the program in routine mode are indeed the unobserved $P(D(0)=1)$ and $ATT(0)$ (Heckman and Smith, 1998), we can view the effect of randomisation on the pool of participants as introducing bias in the experimental estimate for the parameter of interest.

An alternative way to view the causal impact of randomisation on participation decisions is as a potential threat to the *external* validity of the experimental estimate. In this case, the parameter of interest is not the impact of the intervention on the population which would be subject to the intervention in routine mode, but the impact for the sample of study participants, and the issue is the extent to which such conclusions from the trial would generalise to the population which would be exposed to the treatment in routine mode.

Note finally that a number of terms have been used in various disciplines to describe how the behaviour of human participants is affected by being part of a research study. Among these are the:

- *Placebo effect* in medicine, where receiving an inert substance has a positive effect due to the expectation and belief created in the participants;
- *Hawthorne* (or *reactivity*) *effect*, when participants behave differently simply because they know they are being observed and studied in connection with the measured outcomes;
- *John Henry effect*, when the non-treated subjects work harder to compensate for not having been offered the treatment;
- *Novelty effect* leading participants to perform differently because of the novelty of the treatment;

- *Demand characteristics effect* arising from the (possibly perceived) requirement being placed on the study participant to be "good subjects";
- *Pygmalion effect* (Rosenthal and Jacobson, 1968), where expectations held by researchers about the performance of their subjects turn into self-fulfilling prophecies.

In medical trials, blinding (concealing treatment and hence expectations from the patient and if double blind from the experimenter too) controls for these effects by making them equal for both groups. These confounding effects do not however originate from the randomised element of a study, but merely from the investigation *per se*, and as such do not give rise to randomisation bias.[2]

3. Randomisation bias in the clinical literature

It is worth mentioning that deliberate attempts to subvert the intended purpose of randomisation were actually found to occur much more frequently than one would expect by Schulz (1995), who concludes that "Randomized controlled trials appear to annoy human nature – if properly conducted, indeed they should".[3]

Well recognised and more explored issues related to the process of randomisation *per se* that may give rise to randomisation bias and hence affect the validity of conclusions drawn from the results of clinical trials are:

(1) preference effects,
(2) informed consent effects and
(3) participation bias.

In the framework above, the first two factors above my impact on potential outcomes Y_1 and Y_0, while all three may directly affect participation choices D. We discuss these two possibilities in turn.

[2] See McCambridge *et al*. (2014a) for a historical perspective and summary of what they fittingly term "research participation effects", that is effects that arise from the interaction of the research participant with the research process.

[3] Based on anonymous accounts, the author reports attempts ranging "from simple to intricate operations, from transillumination of envelopes to searching for code in the office files of the principal investigator".

3.1 Randomisation can affect potential outcomes

Two channels – individual preferences and informed consent – can interact with randomisation to directly affect potential outcomes Y_1 and Y_0, as clearly recognised by Britton *et al.* (1998) in their extensive review of the validity of clinical RCTs: "randomisation may be misleading where the process of random allocation may affect the effectiveness of the intervention" (p.1). Indeed, as discussed above, when design features like randomisation affect potential outcomes and in particular no-treatment outcomes, the internal validity of the causal inference is undermined.

By definition and in contrast to observational studies, randomised trials prevent individuals (physicians and patients) from expressing their **preferences**. This feature can potentially affect outcomes in two ways: directly via the therapeutic effect of choice and sense of control (akin to a placebo effect) and indirectly via the extent of actual compliance by patients (and/or care by physicians).[4] Let us consider how the control and program group might be affected in this way and, to fix ideas, let us do so in the likely scenario where some patients prefer the new experimental treatment (only available in the trial) to the standard best-practice control treatment.

Those control group patients randomised to their non-preferred allocation may suffer from what Cook and Campbell (1979) termed "resentful demoralisation", which would give rise to randomization bias if it affects their no-treatment outcome Y_0 either directly (through a negative placebo-like effect) or indirectly (through poor or no compliance to the control treatment protocol). These demotivating effects of having a preferred treatment withheld prevent the control group from providing the correct average counterfactual outcome for the program group. This is the most serious form of randomisation bias: these effects purely stem from being part of a randomised trial and lead to (most likely upwards) biased estimates of the impact of the experimental intervention relative to the standard treatment. An alternative to resentful demoralisation is what Cook and Campbell (1979) call "compensatory rivalry" (what in economics is referred to as control group substitution or contamination), which leads disappointed control group members to seek out a different type or intensity of alternative interventions from what they would have done in the absence of the trial.[5]

[4] A recent meta-analysis has found that preferences among patients in musculoskeletal trials are indeed associated with treatment effects (Preference Collaborative Group, 2008).

[5] The qualitative study by McCambridge *et al*. (2014c) documents for instance that disappointment among the control group of a weight loss trial led to movements both toward and away from behaviour change.

In the program group, by contrast, such preference effects have different implications. First, the fact that patients receiving their preferred treatment may comply better than average would not present a problem if interest lies in recovering the average treatment effect for the treated (*ATT*), as compliance amongst the experimental group would be representative of the compliance among patients freely choosing the experimental treatment under routine conditions. It would however likely over-estimate the average treatment effect (*ATE*) in the population of eligible patients (i.e. including those who would not be interested in taking up the new treatment). Similarly, any direct therapeutic effect of patient preference on treatment outcomes would not give rise to randomisation bias, as this therapeutic effect would occur too under normal conditions where patients freely choose their preferred treatment. What it means though is that what is being tested is in fact the *combined* effect of the active treatment and the preference for it; the experimental treatment may thus gain in apparent effectiveness, even if it has no additional physiological benefit. As was the case for compliance effects, the presence of this therapeutic effect of choice would allow for the unbiased estimate of the corresponding *ATT* (in this case, the combined effect for the treated of treatment and its choice), but it will likely overestimate the impact for the population, *ATE*.

Blinding – a design in which patients (and ideally clinicians too, as in double-blinded trials) do not know which treatment arm they have been allocated to – can mitigate these preference effects. Indeed, RCTs that have not used appropriate levels of blinding were found to show on average significantly larger treatment effects than blinded studies (Schulz *et al*., 1995).

It is however often inappropriate or impossible to blind, such as in the case of surgical procedures, when the active participation of patients is required or in the presence of very specific side-effects.[6] An alternative is the "randomised consent" design proposed by Zelen (1979, 1990), in which what is randomly assigned is the seeking of consent to be randomly assigned. Specifically, patients are randomly assigned to either receive the best standard treatment (without being asked) or to be asked if they will take part in the experimental treatment (where those who decline will receive the standard treatment). While this design avoids any potential disappointment among the control group as they do not even know they

[6] In double-blind trials of antidepressant drugs, the vast majority of patients and their clinicians have been found to break the blind early on (most likely due to their monitoring of side effects); see Fisher and Greenberg (1993) and Margraf *et al*. (1991).

are part of a trial, this set-up can only recover an intention-to-treat effect (i.e. the average effect of *offering* the experimental treatment), instead of the *ATT*.[7]

Building on Zelen's work, a number of designs incorporating preference arms have been put forward. Brewin and Bradley's (1989) "partially randomized patient-centred" design allocates patients with a strong preference for one treatment to the treatment they prefer, and patients with no strong preference to the experimental or standard control treatment at random. As both of the randomly allocated groups will be similarly motivated towards the treatment they were allocated to, their comparison is internally valid. It is however informative only about the effect of the experimental treatment on those with no strong preference for it, clearly excluding the policy-relevant group of those who would clearly favour it. Finally, Korn and Baumrind's (1991) "partially randomized clinician-centred" design incorporates doctors' preferences into the process of allocating patients to experimental or standard control treatment. Only those patients where both objective and clinician screenings result in no clear treatment preference are randomly assigned to a clinician preferring the experimental treatment or to a clinician preferring the control treatment. Again while internally valid, the resulting casual parameter has limited external validity, due to the difficulty in characterizing the population to whom the results of the trial could be generalized. So while such modified designs might successfully adjust for preferences, they do raise their own set of methodological and ethical concerns (see the appraisal by Bradley, 1993). Designs intended to detect preference effects have also been proposed, e.g. Rucker (1989).

A systematic review of 32 studies by King *et al*. (2005) found however that while preferences led a considerable proportion of patients to refuse to be randomized, there was less evidence that preferences substantially interfered with the internal validity of the trials.

In addition to preference effects, **informed consent** – a usual requirement of ethics committees approving randomised trials – might give rise to randomisation bias by affecting potential outcomes and the therapeutic response to the treatment. For example, in a trial to analyse the effect of a placebo when administered with or without informed consent, Dahan *et al*. (1986) found that obtaining informed consent spuriously decreased the apparent efficacy of the placebo. In Bergmann *et al*.'s (1994) trial, it decreased the difference between a placebo and active agent, thus undervaluing the treatment.

[7] See Adamson *et al*. (2006) for a review and appraisal of published trials using this method.

Informed consent could potentially affect outcomes also via the educational value of the printed information it provides – either by improving compliance or by conferring a therapeutic benefit in itself. This issue could be particularly serious in behavioural intervention trials, the aim of which is to influence behaviour (e.g. smoking cessation programs). As McCambridge *et al*. (2014b) point out, formally signing a consent form could be perceived as a form of interpersonal declaration of commitment to change, which could potentially affect the behaviour of interest of both the program and control groups. In this case, the outcome would be affected by a feature exclusively linked to randomisation.[8] If under routine conditions patients would be exposed to a different information set or format, the trial would recover a different parameter, in particular, the impact of the experimental treatment packaged with the consent information compared to being exposed to the consent information alone. (This definition allows for a synergistic effect of information provision and behavioural intervention).

Finally, just facing the decision of whether or not to participate in a clinical trial might lead to increased anxiety[9], while the uncertainty inherent in randomization could generate apprehension (Cook and Campbell, 1979). Both of these may in turn affect outcomes in both treatment arms and thus bias experimental estimates in unspecified directions.

3.2 Randomisation can lead to biased sampling of the population

A second set of issues – exclusions, recruitment and patient decisions to participate in an RCT – affect trial participation *D*, thus potentially leading to biased sampling of the population. The concern here is the degree to which patients participating in trials (the "study group") are representative of the full population in need of treatment (the "target group") to whom the results will subsequently be applied.

While the generally low accrual to clinical trials has been widely recognised (e.g. only 2% of eligible patients with breast cancer were participating in US trials in the 1980s), together with the possible impact that selective patient participation might have on research findings, progress in exploring this issue has been particularly slow. A *Lancet* (1992) editorial highlighted the lack of empirical research on the reasons to refuse to participate in trials,

[8] The qualitative study by McCambridge *et al*. (2014c), for instance, found that decisions to participate in a weight loss trial were so closely linked with decisions to change behaviour that for many participants the two were synonymous.

[9] Simes *et al.* (1986) randomly compared uniform total disclosure of all relevant information with an individual approach at the doctor's discretion as methods for obtaining informed consent, finding that the former led to increased anxiety (though this effect was significant only during the first follow-up month).

while the review by Britton *et al.* (1998) found that most RCTs failed to document adequately the characteristics of eligible individuals who did not participate in trials. Rothwell (2005) similarly makes the case for greater consideration of external validity in the design and reporting of RCTs. The subsequent systematic sampling review by Van Spall *et al*. (2007) however highlighted that the RCTs published in major medical journals do not always clearly report exclusion criteria and indeed that only less than half of the reported exclusion criteria were classified as strongly justified within the context of the specific trial.

The process through which patients become involved in trials involves decisions by centres, researchers, clinicians and the patients themselves. Of relevance to the present discussion is that randomisation (together with other features to ensure internal validity) further compounds selective trial participation issues at each step.

The first hurdle is **centre selection**, with the concern that providers who agree to take part may be atypical, e.g. 'centres of excellence'. While not necessarily related to RCTs, a randomised trial is likely to sharpen such centre selection issues.

The second hurdle are the **researchers** who design the trial and draw up the exclusion criteria. While some of these criteria are motivated on medical or ethical grounds, there is a documented tendency, possibly exacerbated by fears of litigation, to rely on blanket exclusion criteria, much wider than in normal clinical practice. Indeed, recent evidence cited in Pressler and Kaizar (2013) spanning trials for alcohol treatments, antidepressants and asthma shows that it is quite common for less than half of the affected population to be eligible to participate in the RCT. Similarly, Bartlett *et al*. (2005) document how older people, women and ethnic minorities are systematically excluded from UK medical research studies. Other real-world patients routinely excluded are those with confounding multiple pathology, severe disease or undergoing other treatments. Other exclusion criteria are based on scientific or administrative grounds and can be directly viewed as a form of randomisation bias. Examples are the use of very narrow criteria to ensure a homogeneous sample and hence a higher likelihood of a statistically significant result from a small-scale RCT; or criteria aimed to exclude patients who are, or are expected to be, difficult to gain informed consent from (e.g. children, the mentally ill, drug users, non native speakers).

The third hurdle are the recruitment and referral decisions by **clinicians**. The decision by the doctor not to offer a patient the option to enter an RCT was found not only to be a major reason for poor accrual to clinical trials, but indeed to reflect doctors' personal discomfort with randomisation *per se* (see e.g. the review by Ellis, 2000). Other common reasons given for not recruiting patients to trials can similarly be viewed as stemming from randomisation:

worries about its impact on the doctor-patient relationship, difficulties with informed consent procedures, aversion to openly discussing uncertainty and finding pragmatic aspects of a trial (time and effort) to be too burdensome (e.g. Taylor *et al*., 1984).

There is also evidence that information on trials is tailored to steer accrual either way. For instance, UK doctors were found to adopt individual methods when providing information and eliciting consent to trials (Jenkins *et al*., 1999). Williams and Zwitter (1994) found that only 62% of surveyed doctors routinely told all of their patients that treatment would be assigned at random and only 58% gave patients information about all of the treatment options. Another example is offered by Senore *et al*. (1999). They estimated the probability that an eligible patient be offered enrolment in a trial of anti-smoking counselling in Italy and found that General Practitioners focused their recruitment on high-risk smokers and on those who had tried to quit.

The final hurdle to trial participation are of course decisions by the **patients** themselves, and indeed the reasons to decline – preference for one particular intervention and unwillingness to be randomly assigned – can be viewed as forms of randomisation bias. Specifically, as mentioned, non-experimental designs allow patients to choose the treatment option they prefer. In an RCT, by contrast, this choice is taken away and patients can only choose whether to receive a probability of receiving a given therapy (indeed, they might only be able to choose whether to volunteer for a blinded trial where they would potentially but not knowingly be receiving a new treatment with unproven benefits and unknown risks). Such a design will thus exclude those with strong treatment preferences (hence not willing to leave the choice to chance) and is likely to create a highly selective pool of trial participants, e.g. those accepting randomisation as a last hope for a cure. In addition to the loss of control, a dislike of randomisation itself is the major reason given by patients for declining trial entry (see e.g. Jenkins and Fallowfield, 2000, and the literature review by Ellis, 2000). Indeed, such recruitment hence selection difficulties are further compounded by the fear of being a 'placebo responder'.[10]

While this trade-off between features (randomisation, blinding, placebo) needed for an unbiased answer to a specific question about an intervention and the need to provide answers of direct relevance to everyday clinical practice has been recognised, "there is at present no clear answer as to how this issue should be resolved" (Bartlett *et al*., 2005, p.103), leading some authors to conclude that "one cannot simply assume that the RCT findings are

[10] For a general framework for conceptualising preferences in RCTs and their effects on decision-making, see Bower *et al*. (2005).

definitive, because the lack of generalizability can sometimes be more serious than the threats to internal validity in the observational studies" (Weisberg *et al*. (2009, p.116).

In their extensive review, Britton *et al*. (1998) find that only a relatively small proportion of relevant patients are included in RCTs and that such participants are not representative of those who are eligible to be included, let alone of the population to whom the results will subsequently be applied. They further highlight how such differences could influence any effect detected, raising concerns that "authors often ignored or discounted clear, statistical evidence that participation bias may have occurred – presumably because they felt it would undermine their findings." (p.34)

Before reviewing how this kind of "participation bias" has been considered in the clinical literature, a conceptual clarification is in order. This paper is explicitly concerned with how randomisation *per se* could give rise to a study sample which is potentially different from the study sample which would have been obtained had the treatment not been evaluated via random assignment. This is the bias introduced by randomisation as such, a bias which is particularly hard to assess as it relates to a counterfactual (and generally unobserved) study population. The literature reviewed in the following, whilst recognising the selected nature of trial participants, is by contrast interested in the wider generalizability issue of moving from the trial study sample to a broader or different 'target population'. There are indeed various factors that can limit the external validity of an RCT (as indeed of an observational study): in addition to the ones outlined above (centre/site selection, eligibility criteria, referral/recruitment or patient self-selection), there is e.g. the issue of efficacy (the effect of treatment under ideal and highly controlled conditions in a research setting) *versus* effectiveness (the effect of the same treatment in real-world, everyday practice) or the wider generalizability issues of scaling up the intervention to the national level, or of offering it to a different population. Keeping thus in mind that while the 'target population' in the following is not necessarily the 'study population which would have arisen in the absence of randomisation', let us now review the various ways in which the non-representativeness of trial participants has been viewed and assessed in the medical literature.

First, there is of course the issue of identifying the 'target population' itself and on this basis of identifying and characterising the eligible participants and the eligible non-participants. To this regard, a number of authors (e.g. Britton *et al.*, 1998, Bartlett *et al*., 2005, and Van Spall *et al*., 2007) have advocated as the immediate priority the improvement of the quality of reporting, as information on eligibility criteria and participation is often simply

lacking even for RCTs published in major medical journals.[11]

Design-wise, there have also been calls to run *practical clinical trials* – large-scale pragmatic RCTs involving participants and treatments that are representative of those in the target population (Tunis *et al*., 2003). This kind of trial, whose aim is generally to test treatment effectiveness, could however at best limit external validity issues arising from eligibility criteria and the referral and recruitment processes, but not from self-selection, as informed consent will always be an integral part of an RCT.

Indeed, instruments such as interviews and questionnaires have been used to explore preference formation and decision-making in clinical RCTs, often probing attitudes towards likely participation in a *hypothetical* RCT and comparing the characteristics of those who would and who would not take part. A recent and extensive example of the latter is Jenkins *et al*. (2010); they however themselves note that there is evidence that a larger number of patients will refuse RCTs in routine clinical practice than in hypothetical surveys.

Studies often just speculate about the potential generalizability of their intervention, mostly assuming no variation in biological response between sub-groups. Clearly, in a world of homogenous treatment impacts any sample selection issue arising from randomisation *per se* (or indeed any other reason) would not pose any generalizability problems (cf. Section 2). The assumption of homogeneous impacts in clinical interventions is an unrealistically strong one, and is refuted e.g. by Britton *et al*. (1998) and Kravitz *et al*. (2004). The latter authors conclude that treatment effect heterogeneity is indeed plausible across a variety of clinical contexts after highlighting how this heterogeneity can result from patient diversity in baseline risk of disease, responsiveness to the treatment, vulnerability to the adverse effects of the treatment and utilities (i.e. patients' values and preferences) for different outcomes.[12]

More comprehensive frameworks for the evaluation of external validity have been put forward based on qualitative studies such as in integral process evaluations and checklists. Glasgow *et al*. (1999) introduced the RE-AIM model for evaluating public health interventions across the five dimensions of Reach, Efficacy, Adoption, Implementation and Maintenance. This paper has been followed by over 100 publications on RE-AIM by a variety of authors in diverse medical fields. In subsequent work, Green and Glasgow (2006) have proposed questions, guides, criteria and procedures to determine the generalizability of trials results, while e.g. Bornhöft *et al*. (2006) have independently defined a number of criteria to

[11] Implementation of the CONSORT statement will only partially address this issue as it does not include the requirement to report the characteristics of those included and excluded.

[12] Willke *et al*. (2012) offer a primer on methods to explore heterogeneity of treatment effects drawing from the biomedical, statistical, epidemiological and econometrics literature.

qualitatively assess a trial's external validity and translate them into a comprehensive checklist in questionnaire format. More recently still, centre selection has been explored by Gheorghe *et al*. (2013) using a mixed methods approach consisting of a systematic review and meta-summary of centre selection criteria reported in RCT protocols, an online survey and focus groups.

In terms of quantitative approaches, evidence-synthesis techniques have been used by some studies. For example, based on meta-analysis, Dhruva and Redberg (2008) find that the demographics of the Medicare population differ significantly from the cardiovascular clinical trial participants used to inform Medicare coverage decisions. After investigating the exclusion of women, older people and ethnic minorities from trials of statins and nonsteroidal anti-inflammatory drugs, Bartlett *et al*. (2005) explore the use of fixed-effects meta-analysis and random-effects meta-regression to estimate the level of relative effectiveness of the drugs in the excluded groups and to assess whether they differ greatly from those in the well-represented groups.

Extending the logic of meta-analysis, researchers have investigated methods for combining results from studies of fundamentally different designs using cross-design synthesis methods introduced by US Government Accountability Office (1992). This broad class of methods combines results from randomized and non-randomized studies in a single analysis that aims to capitalise on the high internal validity of RCTs and the high external validity of observational studies. These approaches typically require a large number of studies and need to meet several challenges in order to first adjust and then combine such diverse studies. Bayesian methods can be used to e.g. model various parameters from each study, include study characteristics and incorporate expert judgment on the relative merits of different types of evidence.

More recent contributions exploit observational data in simpler and more direct ways to assess the issue of the external validity of trial results for a suitably defined target population. For instance, Pressler and Kaizar (2013) rely on observational data alone to assess what they term 'generalizability bias' by splitting up the observational sample into those who would and would not be eligible to participate in the trial and estimating the impact separately for the two groups. Greenhouse *et al*. (2008) offer a case study to judge the generalizability of RCTs by comparing the characteristics and outcomes of the depressed adolescents who participated in trials to the characteristics and the adjusted outcomes of depressed adolescents in the United States as estimated from observational data.

Theoretical and modelling work in the clinical literature has also advanced our understanding of subtle issues relating to bias in trial impact estimates due to selection issues. Weisberg *et al*. (2009) for instance propose a simplified counterfactual model with a binary outcome to illustrate that if *a priori* high-risk (low-risk) patients are more likely to be excluded, the risk ratio from the trial will be upward (downward) biased for the true one in the target population, to the point that it might even be of a different sign. Frangakis (2009) stresses that differences between trial participants and the target population can also translate into variables measured after the treatment – in particular any mediator of treatment effects such take-up or compliance – and provides an example framework for calibrating treatment effects using such post-treatment measures from an RCT to a study target population.

4. Conclusions

Individual preferences and informed consent may not only affect the pool of trial participants but also interact with randomisation to directly affect potential outcomes. Participants in clinical RCTs have also been found to be generally quite unrepresentative of the population to whom the results will subsequently be applied. In fact, this highly selected group has had to go through a series of hurdles of centre selection, eligibility, invitation by clinicians and decision to participate, and indeed randomisation and related features to ensure internal validity (such as blinding and placebo) can further compound selective participation issues at each step.

The assessment of medical treatments via clinical trials and the evaluation of microeconomic interventions via social experiments face in principle many common threats arising from randomisation *per se*, in particular those affecting the pool of participants. Recent examples from the econometrics literature include Allcott (2012) and Sianesi (2010 and 2014). The former considers "site selection bias", akin to the centre selection issue in the medical literature. This bias arises when the probability that a program is implemented is correlated with its impacts. While not necessarily related to RCTs, the prospect of a randomised experiment may exacerbate site selection and thus raise serious issues for the wider generalisability question of drawing inference about a program's impact at full scale.[13] Sianesi (2014) considers a large-scale social experiment in the UK, the ERA demonstration, in which the process of randomly allocating individuals has modified who participates in the

[13] A similar example is Hotz *et al*. (2005), who consider the extrapolation of experimental impacts of a job-training program to other sites.

experiment and as a consequence has significantly altered the treatment effect being recovered. In this trial, randomisation bias has led to the exclusion of the "diverted customers" by caseworkers, akin to those selected out by clinicians, and of the "formal refusers", akin to those who would not refuse the treatment but refuse to be randomly assigned. Sianesi (2014) recovers estimates of the extent of randomisation bias by relying on the assumption of no residual selection into the trial based on unobserved idiosyncratic impact components. Sianesi (2010) further extends the methodological framework to allow selection into trial participation to depend on outcome-relevant unobservables by considering econometric approaches which build on the classical sample selection model of Heckman (1979). She also outlines a number of model specification tests.

In view of these common issues, sharing methodological insights and tools could thus prove very beneficial to both disciplines in tackling this rather neglected but highly fascinating and policy relevant topic.

References

Adamson, J., Cockayne, S., Puffer, S. and Torgerson, D.J. (2006), "Review of randomised trials using the post-randomised consent (Zelen's) design", *Contemporary Clinical Trials*, 27, 305-319.

Allcott, H. (2012), "Site selection bias in program evaluation", NBER Working Paper No.18373.

Bartlett, C., Doyal, L., Ebrahim, S., Davey, P., Bachmann, M., Egger, M. and Dieppe, P. (2005), "The causes and effects of socio-demographic exclusions from clinical trials", *Health Technology Assessment*, 9, 1-168.

Bergmann, J.-F., Chassany, O., Gandiol, J., Deblois, P., Kanis, J.A., Segrestaa, J.M., Caulin, C. and Dahan, R. (1994), "A randomised clinical trial of the effect of informed consent on the analgesic activity of placebo and naproxen in cancer pain", *Clinical Trials and Meta-analysis*, 29, 41-47.

Bornhöft, G., Maxion-Bergemann, S., Wolf, U., Kienle, G.S., Michalsen, A., Vollmar, H.C., Gilbertson, S. and Matthiessen, P.F. (2006), "Checklist for the qualitative evaluation of clinical studies with particular focus on external validity and model validity", *BMC Medical Research Methodology*, 11, 6-56.

Bower, P., King, M., Nazareth, I., Lampe, F. and Sibbald, B. (2005), "Patient preferences in randomised controlled trials: conceptual framework and implications for research", *Social Science & Medicine*, 61, 685–695

Bradley, C. (1993), "Designing medical and educational intervention studies: A review of some alternatives to conventional randomized controlled trials", *Diabetes Care*, 16, 509-518.

Brewin, C.R. and Bradley, C. (1989), "Patient preferences and randomized clinical trials", *British Medical Journal*, 299, 313-15.

Britton, A., McPherson, K., McKee, M., Sanderson, C., Black, N. and Bain, C. (1998), "Choosing between randomised and non-randomised studies: A systematic review", *Health Technology Assessment*, 2, 1-124.

Cook, T.D. and Campbell, D.T. (1979), *Quasi-experimentation: Design and* Analysis *Issues for Field Settings*, Chicago, Rand McNally.

Dahan R., Caulin C., Figea L., Kanis J.A., Caulin F., Segrestaa J.M. (1986), "Does informed consent influence therapeutic outcome? A clinical trial of the hypnotic activity of placebo in patients admitted to hospital", *British Medical Journal*, 293, 363-4.

Dhruva, S.S. and Redberg, R.F. (2008), "Variations between clinical trial participants and Medicare beneficiaries in evidence used for Medicare national coverage decisions", *Archives of Internal Medicine*, 168, 136-40.

Ellis, P.M. (2000), "Attitudes towards and participation in randomised clinical trials in oncology: A review of the literature", *Annals of Oncology*, 11, 939-945.

Fisher, S., Greenberg, R.P. (1993), "How sound is the double-blind design for evaluating psychotropic drugs?", *Journal of Nervous and Mental Disease,*181, 345–350.

Frangakis, C. (2009), "The calibration of treatment effects from clinical trials to target populations", *Clinicals Trials*, 6**,** 136–140.

Green, L.W. and Glasgow, R.E. (2006), "Evaluating the relevance, generalization, and applicability of research: Issues in external validation and translation methodology", *Evaluation & the Health Professions*, 29, 126–153.

Greenhouse, J.B., Kaizar, E.E., Kelleher, K., Seltman, H. and Garnder, W. (2008), "Generalizing from clinical trial data: A case study. The risk of suicidality among pediatric antidepressant users", *Statistics in Medicine*, 27, 1801–1813.

Glasgow, R. E., Vogt, T. M. and Boles, S. M. (1999), "Evaluating the public health impact of health promotion interventions: the RE-AIM framework", *American Journal of Public Health*, 89, 1322–1327.

Heckman, J.J. (1979), "Sample Selection Bias as a Specification Error", *Econometrica*, 47, 153-161.

Heckman, J.J. (1992), "Randomization and social policy evaluation", in C. Manski and I. Garfinkel (eds.), *Evaluating welfare and training programs*, Harvard University Press, 201-230.

Heckman, J.J. and Smith, J. (1998), "Evaluating the welfare state," in S. Strom (ed.), *Econometrics and Economics in the 20th Century*, Cambridge University Press, New York.

Hotz, V.J., Imbens, G.W. and Mortimer, J.H. (2005), "Predicting the efficacy of future training programs using past experiences at other locations", *Journal of Econometrics*, 125, 241-270.

Imbens, G.W. and Angrist, J.D. (1994), "Identification and estimation of local average treatment effects", *Econometrica*, 62, 446-475.

Jenkins, V. and Fallowfield, L. (2000), "Reasons for accepting or declining to participate in randomized clinical trials for cancer therapy", *British Journal of Cancer*, 82, 1783-1788.

Jenkins, V., Fallowfield, L., Souhami, A. and Sawtell, M. (1999), "How do doctors explain randomised clinical trials to their patients?", *European Journal of Cancer*, 35:1187-1193.

Jenkins, V., Farewell, D., Batt, L., Maughan, T., Branston, L., Langridge, C., Parlour, L., Farewell, V. and Fallowfield, L. (2010), "The attitudes of 1066 patients with cancer towards participation in randomised clinical trials", *British Journal of Cancer*, 103, 1801-1807.

King, M., Nazareth, I., Lampe, F., Bower, P., Chandler, M. and Morou M. (2005), "Impact of participant and physician intervention preferences on randomized trials: A systematic review", *The Journal of the American Medical Association*, 293, 1089-1099.

Korn, E.L. and Baumrind, S. (1991), "Randomized clinical trials with clinician-preferred treatment", *Lancet*, 337, 149-52.

Kravitz, R. L., Duan, N. and Braslow, J. (2004), "Evidence-Based Medicine, Heterogeneity of Treatment Effects, and the Trouble with Averages", *The Milbank Quarterly*, 82, 661–687.

Lancet (1992), "Volunteering for research", Editorial, *Lancet*, 340, 823-4.

Margraf, J., Ehlers, A., Roth, W.T., Clark, D.B., Sheikh, J., Agras, W.S. and Taylor, C.B. (1991), "How "blind" are double-blind studies?", *Journal of Consulting and Clinical Psychology*, 59, 184-187.

McCambridge, J., Kypri, K. and Elbourne, D. (2014a), "Research participation effects: A skeleton in the methodological cupboard", *Journal of Clinical Epidemiology*, 67, 845-849.

McCambridge, J., Kypri, K. and Elbourne, D. (2014b), "In randomization we trust? There are overlooked problems in experimenting with people in behavioral intervention trials", *Journal of Clinical Epidemiology*, 67, 247-253.

McCambridge, J., Sorhaindo, A., Quirk, A., & Nanchahal, K. (2014c), "Patient preferences and performance bias in a weight loss trial with a usual care arm", *Patient Education and Counseling*, *95*, 243–247.

Pressler, T.T. and Kaizar, E.E. (2013), "The use of propensity scores and observational data to estimate randomized controlled trial generalizability bias", *Statistics in Medicine*, 32, 3552-68.

Preference Collaborative Group (2008), "Patients' preferences within randomised trials: Systematic review and patient level meta-analysis", *British Medical Journal*, 337, 1-8.

Rosenthal, R. and Jacobson, L. (1968), *Pygmalion in the classroom: Teacher expectation and pupils' intellectual development*. New York: Holt, Rinehart & Winston.

Rothwell, P. M. (2005), "External validity of randomised controlled trials: To whom do the results of this trial apply?" *The Lancet*, 365, 82–93.

Rucker, G. (1989), "A two-stage trial design for testing treatment, self-selection and treatment preference effects", *Statistics in Medicine*, 8, 477-85.

Schulz, K.F. (1995), "Subverting randomization in controlled trials", *Journal of the American Medical Association*, 274, 1456-1458.

Schulz, K.F., Chalmers, I., Hayes, R. and Altman, D.G. (1995), "Empirical evidence of bias: dimensions of methodological quality associated with estimates of treatment effects in controlled trials", *Journal of the American Medical Association*, 273, 408-412.

Senore, C., Battista, R., Ponti, A., Segnan, N., Shapiro, S., Rosso, S. and Aimar, D. (1999), "Comparing participants and nonparticipants in a smoking cessation trial: Selection factors associated with general practitioner recruitment activity", *Journal of Clinical Epidemiology*, 52, 83-89.

Sianesi, B. (2010), "Supplementary Technical Appendix for 'Non-participation in the Employment Retention and Advancement study: Implications for the experimental first-year impact estimates'", linked to Department for Work and Pensions Technical Paper No. 77.

Sianesi, B. (2014), "Dealing with randomisation bias in a social experiment: The case of ERA", IFS Working Paper No.W14/10.

Simes, R.J., Tattersall, M.H.N, Coates, A.S., Raghavan, D., Solomon, H.J. and Smartt, H. (1986), "Randomised comparison of procedures for obtaining informed consent in clinical trials of treatment for cancer", *British Medical Journal*, 293, 1065-8.

Taylor, K.M., Margolese, R.G. and Soskolne, C.L. (1984), "Physicians' reasons for not entering eligible patients in a randomized clinical trial of surgery for breast cancer", *New England Journal of Medicine*, 310, 1363-7.

Tunis, S.R., Stryer, D.B. and Clancy, C.M. (2003), "Practical clinical trials: Increasing the value of clinical research for decision making in clinical and health policy", *The Journal of the American Medical Association*, 290, 1624-32.

US Government Accountability Office (GAO) (1992), "Cross Design Synthesis: A New Strategy for Medical Effectiveness Research", PEMD-92-18.

Van Spall, H., Toren, A., Kiss, A. and Fowler, R.A. (2007), "Eligibility criteria of randomized controlled trials published in high-impact general medical journals: A systematic sampling review", *Journal of the American Medical Association,* 297, 1233-1240.

Weisberg, H.I., Hayden, V.C. and Pontes, V.P. (2009), "Selection criteria and generalizability within the counterfactual framework: Explaining the paradox of antidepressant-induced suicidality?", *Clinical Trials*, 6, 109–118.

Williams, C.J. and Zwitter, M. (1994), "Informed consent in European multi-centre randomised clinical trials: Are patients really informed? *European Journal of Cancer*, 30A, 907-10.

Willke, R.J., Zheng, Z., Subedi, P., Althin, R. and Mullins, C.D. (2012), "From concepts, theory, and evidence of heterogeneity of treatment effects to methodological approaches: A primer", *BMC Medical Research Methodology*, 12, 1-12.

Zelen, M. (1979), "A new design for randomized clinical trials", *New England Journal of Medicine*, 300, 1242-45.

Zelen M. (1990), "Randomized consent designs for clinical trials: An update", *Statistics in Medicine*, 9, 654-656.