# USING OBSERVATIONAL VS RANDOMIZED CONTROLLED TRIAL DATA TO LEARN ABOUT TREATMENT EFFECTS

BRENDAN KLINE AND ELIE TAMER

NORTHWESTERN UNIVERSITY

Abstract. Randomized controlled trials (RCTs) are routinely used in medicine and are becoming more popular in economics. Data from RCTs are used to learn about treatment effects of interest. This paper studies what one can learn about the average treatment response (ATR) and average treatment effect (ATE) from RCT data under various assumptions and compares that to using observational data. We find that data from an RCT need not point identify the ATR or ATE because of selection into an RCT, as subjects are not randomly assigned from the population of interest to participate in the RCT. This problem relating to external validity is the primary problem we study. So, assuming internal validity of the RCT, we study the identified features of these treatment effects under a variety of weak assumptions such as: mean independence of response from participation, an instrumental variable assumption, or that there is a linear effect of participation on response. In particular we provide assumptions sufficient to point identify the ATR or the ATE from RCT data and also shed light on when the sign of the ATE can be identified. We then characterize assumptions under which RCT data provide more information than observational data.

Keywords: randomized controlled trials, experiments, treatment effect, identification

*Date*: February 1, 2011.

$*$ Department of Economics, Northwestern University, 2001 Sheridan Road, Evanston, IL 60208, USA. We thank Chuck Manski for useful comments. Financial support from the NSF is gratefully acknowledged. Any errors are ours.

# 1. Introduction

This paper is concerned with identification of treatment response in a population of subjects each of whom is characterized by a response function $y_i(d) : d \in \mathcal{D}$ where $d$ is a treatment that belongs to a finite set of mutually exclusive treatments $\mathcal{D}$. We are interested in functionals of the joint distribution of $\{y_i(.)\}$ and focus on the Average Treatment Response (ATR) of treatment $d$ which is

$$ATR(d) \equiv E(y_i(d))$$

and the Average Treatment Effect (ATE) of treatment $d'$ versus treatment $d$ which is

$$ATE(d', d) \equiv E(y_i(d')) - E(y_i(d)).$$

The problem is that we do not observe data on $y_i(d)$ for a subject $i$ at all treatments $d$. Rather, each subject is observed to experience only one treatment. In general, observational data is subject to selection bias (or endogeneity) when response is not mean independent of realized treatment. This is a ubiquitous problem in economics, but also in many other settings like medicine. For example, there is a concern that treatment choice is correlated with unobservables that are also correlated with outcome. The standard example of this in economics is that achieved education is correlated with various unobservables like ability that are also correlated with wages. Consequently, the average wage of people with a given level of educational achievement is not equal to the counterfactual average wage of all people were they all to achieve that level of education.

One possible solution to this problem is a randomized controlled trial (RCT). These have been used since roughly the 1950s in medicine, although there are antecedents dating back perhaps as far as the middle of the 18th century (e.g., see Pocock (1983)), and have since come to be commonly known as the "gold standard" for statistical evidence. They are required, for example, in any application to the FDA to market a new drug. RCTs are also used increasingly often in economics, for example in labor economics, and especially more recently in development economics. See for example Banerjee and Duflo (2009) for a review of the use of experiments in development economics. Effectively the defining characteristic of an RCT is that treatments are assigned randomly to participants, overcoming the possible selection bias problem of observational data relating to treatment selection.

This paper's contribution is to analyze what can be learned about the ATR and ATE with data from an RCT under certain maintained assumptions and then to contrast that with what can be learned with observational data. The main issue we study is *selection into an RCT*, an issue that depends on the way in which an experiment is ran, the way that subjects decide whether to participate in an RCT, and the population of interest. This selection problem threatens the generalizability of the results from an RCT beyond the sample of participants in the study; this is commonly referred to as *external validity.*

The selection into an RCT problem arises because, in general, participation in an RCT is a decision in much the same way that treatment selection is a decision in observational data. Often times the decision to participate in an RCT reflects an even more deliberate thought process than the decision of which treatment to select in observational data because of ethical and legal requirements on experiments involving human subjects. Indeed, the Declaration of Helsinki[1] states that in medical research "the physician or another appropriately qualified individual must then seek the potential subject's freely-given informed consent, preferably in

[1] Available online at http://www.wma.net/en/30publications/10policies/b3/index.html.

writing" and that "[t]he physician may combine medical research with medical care only to the extent that the research is justified by its potential preventive, diagnostic or therapeutic value and if the physician has good reason to believe that participation in the research study will not adversely affect the health of the patients who serve as research subjects."

The ethics of medical RCTs is not settled but one standard view is that "[p]hysicians who are convinced that one treatment is better than another for a particular patient cannot ethically choose at random which treatment to give, they must do what they think best for the patient. For this reason, physicians who feel they already know the answer cannot enter their patients into a trial" and "[a]n ethical physician must do what is best for his or her patients. She cannot participate in a controlled trial if she is certain that one arm is superior to the others and that some of her patients will receive an inferior treatment by participating in the trial" (from the debate in Weijer, Shapiro, Glass, and Enkin (2000); note that one of these quotes is from one side of the debate and the other quote is from the other side). This suggests that, especially on ethical grounds, participation in a medical RCT is statistically related to the outcome. Similarly, the same optimizing behavior that suggests endogeneity in economic models of treatment-response suggest endogeneity of selection into RCTs. Furthermore, although there does not yet exist a large literature on the ethics of RCTs in economics, it seems that especially in the case of development economics in which the subjects are potentially "a disadvantaged or vulnerable population or community" in the language of the Declaration of Helsinki, similar ethical concerns apply to RCTs in economics. So, generally, participation rates in an RCT from among a population of interest can be much lower than in an observational study using the same population of interest, and this participation is selective.

We are interested in external validity, meaning the ability of data to identify the ATR and ATE defined on the population of interest, and our results do not apply to experiments in which one is interested in the ATR and ATE for just the sample of people in the experiment. Indeed, in our assumptions we stipulate that the RCT is "ideal" in the sense that there are no threats to internal validity; that is, the RCT is assumed to point identify the ATR and ATE for the subpopulation actually participating in the RCT. So, issues of compliance, blinding vs nonblinding, and matters relating to the conduct of the experiment are assumed away, not because we do not believe these issues are important, but rather because we intend to focus on the issue of external validity.

Our approach to identification is similar to the one used in Manski (1996) in his analysis of learning about treatment effects from experimental data, in the sense that we study what one can learn about these objects of interest under weak assumptions using a partial identification approach. Manski (1996) studies the basics of treatment compliance (e.g., internal validity) and external validity. We focus exclusively on the details of who participates and assume full compliance. Our setup of the problem is different, as we assume that participation in an RCT is a two step procedure, similar to the stylized model presented by, for example, Gross, Mallory, Heiat, and Krumholz (2002) (henceforth GMHK). The first step is the *invitation step* where researchers invite a group of subjects from the population. This we assume is observed with the key condition that response is mean independent of whether a subject is invited. The second step is the *participation step* where invitees decide whether to participate. We only observe the participants. This is the source of the selection problem. His main model of external validity is basically what we consider as model 2, where RCT data is effectively completely uninformative. In contrast, in our model 1 RCT data can be informative. Manski

(1996) briefly considers a sort of one stage version of our two stage model 1, but as we explain in our discussion of our model, we view the two stage formulation as an important feature of the model. In addition, an important feature of this paper is the comparison of what one can learn from RCT vs observational data. Finally, we consider additional assumptions that tighten the identified set. So, we view both papers as containing complementary results on what one can learn from experimental data under different sets of assumptions. Similar papers in economics that dealt with learning from experimental data and comparing that to what one can learn from observational data are Heckman (1992) and Heckman (1996).

In section 2, we introduce the basic model and derive our identification results. We find that data from an RCT need not point identify the ATR or ATE. We characterize assumptions sufficient for an RCT to point identify either the ATR or the ATE in the population of interest. One such assumption is that response is mean independent of participation, but others are that there are instruments for participation or that there is a linear effect of participation on response. We also consider the identification of the sign of the ATE, which is the object of interest when the experimenter wants to learn the "best" treatment and is abstracting away from issues like different costs of different treatments. Section 3 considers a slightly different model of an RCT that results in completely uninformative identification. Section 4 compares identification with data from an RCT to observational data. We characterize when RCT data results in a narrower identified set than does observational data. Section 5 illustrates our identification results using an analysis of RCTs in medicine. Finally section 6 concludes with some suggestions for reporting and interpreting the results of RCTs. We illustrate our theoretical discussion throughout the paper by real-world examples from RCTs in medicine and economics.

## 2. Model 1: an RCT with observed invitation

The finite set of possible treatments is $\mathcal{D}$, and each subject $i$ in the population of interest has a response function $y_i(\cdot) : \mathcal{D} \to \mathbb{R}$. The stylized randomized controlled trial (RCT) we study is conducted on some subpopulation of the population of interest, and is composed of three decisions made at the subject level. The first decision is the decision by the experimenter of whether to invite the subject to participate in the RCT. The variable indicating an invitation is $I_i$. We assume that invitation is observed. The second decision is the decision by the invited subject of whether to participate in the RCT. The variable indicating participation is $P_i$. We assume that participation is observed. By convention, if $I_i = 0$ we set $P_i = 0$. Finally, the third decision is the randomized assignment of a treatment to each subject participating in the RCT. The variable indicating assigned treatment is $D_i$. We assume that treatment is observed; further, in order to focus on "pre-assignment" biases we abstract from many important issues that might threaten internal validity of an RCT and in particular abstract away from issues related to treatment compliance and assume that all subjects comply with their assigned treatment. The actual outcome of subject $i$ is $y_i$. Since we consider the case of perfect compliance we have $y_i = y_i(d)$ when $D_i = d$. We assume that there are known finite bounds on the possible outcomes, so that necessarily $y_i(d) \in [m, M]$ for all treatments $d$.

Note that for any RCT called 1 the information is effectively the same as for an RCT called 2 in which all subjects in the population of interest are considered invited, but in which subjects who were not invited in RCT 1 are considered to not participate in RCT 2. We separate the two decisions to make clear the meaning of our mean independence assumptions, since we view it as fruitful to distinguish between the invitation decision that is under the control of the experimenter from the participation decision that is not under the control of the experimenter, even though they collectively determine which subjects "actually" participate in the RCT. This two step model of an RCT is similar to the stylized model presented by, for example, GMHK.

We focus first on the case of identifying $E(y_i(d))$, the average response to treatment $d$. We abstract away from including regressors in the analyses for simplicity. In everything that follows we could allow non-parametric conditioning on regressors. By the law of iterated expectation we have that

$$E(y_i(d)) = \underbrace{E(y_i(d)|I_i = 1)}_{(1)} \underbrace{P(I_i = 1)}_{(2)} + \underbrace{E(y_i(d)|I_i = 0)}_{(3)} \underbrace{P(I_i = 0)}_{=1-(2)}.$$

In other words, the average response to treatment $d$ can be decomposed as the weighted average of the average treatment responses for subjects invited, (1), and not invited, (3), to the RCT where the weight is the probability of invitation, (2). The first substantiative assumption we make is that response is mean independent of invitation to the experiment.

**Assumption 2.1.** *Assume that $E(y_i(d)|I_i = 1) = E(y_i(d)) = E(y_i(d)|I_i = 0)$ for all treatments $d \in \mathcal{D}$. Also, $P(I_i = 1) > 0$.*

Under this assumption it is enough to study the identification of $E(y_i(d)|I_i = 1)$. Note that many experiments study the response to an experimental treatment that is not available to subjects not invited to participate in the RCT. Any possible data alone is completely uninformative about $E(y_i(d)|I_i = 0)$ when $d$ is such an experimental treatment, since no subject not participating in the experiment experiences treatment $d$. In our model we maintain this condition in establishing sharpness of the bounds; in principle, other data could be combined with data from the RCT to identify $E(y_i(d)|I_i = 0)$.

The credibility of assumption 2.1 depends on the relationship between the experimental design and the population of interest. It is always logically possible to define the population of interest to be exactly the invited population, in which case $I_i = 1$ for all subjects according to our model. However, in many cases there is population of interest defined beyond the limited scope of a particular RCT. For example, consider the case of heart failure. This disease tends to affect the elderly, affects men and women in roughly equal proportion, and has a significant incidence among non-white people. The most natural population of interest for a treatment for heart failure is apparently the population of people with heart failure. However, Heiat, Gross, and Krumholz (2002) find that RCTs of heart failure treatments often explicitly do not invite the elderly and/or women. The result of this (and, possibly, also participation rates captured by $P_i$ in our model) is that the average age of participants is about 60, even though the average age for patients with heart failure is about 80, only about 20% of participants are women even though about 50% of patients with heart failure are women, and only about 15% of participants are non-white even though about 30% of patients with heart failure are non-white.[2] Moreover, an overwhelming number of RCTs do not invite participation by subjects with certain co-morbidities (i.e., based on left ventricular

[2] Note that it could be that these RCTs have low participation by these groups while not "explicitly" rejecting the elderly, women, or minorities if, for example, they are run in settings with a non-representative sample of patients. In particular, note that no RCT explicitly rejects non-whites, but still their participation rate is lower than the fraction of heart failure patients who are non-white. We are not interested in the precise details of heart failure RCTs, so do not speculate why this has happened. In our model this amounts to not "inviting" these subjects, raising concerns about the credibility of the mean independence assumption. Alternatively, it could be that certain groups have lower propensities to participate in the experiments. This is accounted in our model with the variable $P_i$.

ejection fraction and renal insufficiency). The result is that the subpopulation that tends to be invited to participate in RCTs is not necessarily representative of the population of patients with heart failure. In this case, the mean independence assumption seems not credible. See also Van Spall, Toren, Kiss, and Fowler (2007) for a more general study of invitation to medical RCTs.

The case for economics RCTs is similar. Some RCTs in economics by their very nature effectively invite everyone in the population of interest. Other RCTs necessarily invite only a subpopulation of the population of interest to participate. For example, many experiments in development economics either implicitly or explicitly have as the population of interest something like "developing countries" or "people living in developing countries" but are able to run the experiment only in a geographically restricted area. Related to this point, Banerjee and Duflo (2009) summarize that "experiments in development economics have often been carried out by randomizing over a set of locations or cluster (villages, neighborhoods, schools) where the implementing organization is relatively confident of being able to implement." This may imply that response is not mean independent of invitation, depending on why implementation is possible in some areas but not in others. In general, the credibility of assumption 2.1 depends on how the subjects are invited, and in particular whether they are invited in a way that is correlated with their response. This is an issue that requires knowledge of the details of a particular RCT, so no general discussion is possible here. See also the closely related discussion of "environmental dependence" in Banerjee and Duflo (2009).

If invitation does not satisfy assumption 2.1 it can be useful to "re-define" invitation in the model in such a way that invitation does satisfy assumption 2.1, and that all selection is accounted for in the participation variable $P_i$. The identification power of $I_i$ and associated assumption 2.1 is to maintain that some reasonably sized subpopulation (e.g., the population invited to participate) can be taken to be representative of the population of interest. So, effectively assumption 2.1 maintains that the subjects who are invited to participate are a random sample from the population of interest; the possibility that participation among the invited is non-random is the focus of this paper. Without this assumption, in a model that focuses on selective participation it is difficult to rule out the possibility that those who are involved in the RCT are arbitrarily un-representative of the population of interest. The result would be, when the number of subjects involved in the RCT is very small compared to the size of the population of interest, an almost completely uninformative identified set. Our model 2 in section 3 considers a similar issue.

Regardless of whether we maintain that response is mean independent of invitation to the experiment, we can use the law of iterated expectation a second time to conclude that

$$E(y_i(d)|I_i = 1) = \underbrace{E(y_i(d)|P_i = 1, I_i = 1)}_{(1)} \underbrace{P(P_i = 1|I_i = 1)}_{(2)} + \underbrace{E(y_i(d)|P_i = 0, I_i = 1)}_{(3)} \underbrace{P(P_i = 0|I_i = 1)}_{=1-(2)}.$$

In other words, the average response to treatment $d$ among those invited to participate can be decomposed as the weighted average of the average treatment response for subjects participating, (1), and not participating, (3), in the RCT where the weight is the probability of participation given invitation, (2). As before, the data alone is completely uninformative about $E(y_i(d)|P_i = 0, I_i = 1)$ when $d$ is an experimental treatment, since no subject not participating in the experiment experiences treatment $d$. Also as before, in our model we maintain this condition in establishing sharpness of the bounds. We assume that the RCT is "ideal" in the sense that it point identifies $E(y_i(d)|P_i = 1, I_i = 1)$. A sufficient condition for this is the usual mean independence of response from treatment assignment (and perfect compliance) that usually defines an RCT.

**Assumption 2.2.** *Assume that* $E(y_i(d)|P_i = 1, I_i = 1) = E(y_i(d)|D_i = d, P_i = 1, I_i = 1)$*. Assume further that* $P(P_i = 1|I_i = 1) > 0$ *and* $P(D_i = d|P_i = 1, I_i = 1) > 0$ *for all treatments* $d \in \mathcal{D}$*.*

Under this assumption we have that $E(y_i(d)|P_i = 1, I_i = 1) = E(y_i|D_i = d, P_i = 1, I_i = 1)$. In other words, the average response to treatment $d$ is point identified *for subjects actually participating in the experiment.* These two assumptions seem to exhaust the assumptions that we can credibly make on a stylized RCT. The resulting identified set for $E(y_i(d))$ is given by the following theorem.

**Theorem 2.1.** *Under assumptions 2.1 and 2.2, the sharp identified set for* $E(y_i(d))$ *is that* $E(y_i(d)) \in E(y_i|D_i = d, P_i = 1, I_i = 1)P(P_i = 1|I_i = 1) + [m, M]P(P_i = 0|I_i = 1)$*. Further, the sharp identified set for* $\{E(y_i(d))\}_{d\in\mathcal{D}}$ *is the Cartesian product of these sets.*

*Proof.* The previous discussion establishes these bounds. Sharpness is obtained by considering the response functions

$$y_i(d) = \begin{cases} y_i & \text{if } D_i = d, P_i = 1, I_i = 1 \\ E(y_i|D_i = d, P_i = 1, I_i = 1) & \text{if } D_i \neq d, P_i = 1, I_i = 1 \\ [m, M] & \text{if } P_i = 0, I_i = 1 \\ E(y_i(d)|I_i = 1) & \text{if } I_i = 0. \end{cases}$$

These response functions are consistent with the data by the first line of the definition, are consistent with assumption 2.1 by the fourth line, and are consistent with assumption

2.2 by the second line. They also obviously achieve any point in the identified set by the third line. $\square$

**Corollary 2.1.** *Under the same conditions, the sharp identified set for* $ATE(d',d') \equiv E(y_i(d')-y_i(d))$ *is* $ATE(d',d') \in (E(y_i|D_i=d',P_i=1,I_i=1)-E(y_i|D_i=d,P_i=1,I_i=1))\,P(P_i=1|I_i=1)+[m-M,M-m]P(P_i=0|I_i=1)$.

This corollary follows from the theorem since the identified set for $E(y_i(d')) \times E(y_i(d))$ is the Cartesian product of the marginal identified sets, since there are not restrictions across treatments. The next corollary considers identification of the sign of the ATE, which may be of independent importance in some cases. This is the case when the experimenter wants to learn the "best" treatment and is abstracting away from issues like different costs of different treatments, which would lead to the magnitude of the ATE mattering.

**Corollary 2.2.** *Define the experimental ATE as* $ATE_{exp}(d',d) \equiv E(y_i|D_i=d',P_i=1,I_i=1)-E(y_i|D_i=d,P_i=1,I_i=1)$. *Under the same conditions,* $ATE(d',d)$ *is point identified to be positive (or non-negative, resp.) if* $ATE_{exp}(d',d)P(P_i=1|I_i=1)+(m-M)P(P_i=0|I_i=1)>(\geq)0$, *to be negative (or non-positive, resp.) if* $ATE_{exp}(d',d)P(P_i=1|I_i=1)+(M-m)P(P_i=0|I_i=1)<(\leq)0$, *and is not identified and can be positive, negative, or zero if* $ATE_{exp}(d',d)P(P_i=1|I_i=1)+(m-M)P(P_i=0|I_i=1)<0$ *and* $ATE_{exp}(d',d)P(P_i=1|I_i=1)+(M-m)P(P_i=0|I_i=1)>0$.

The first key conclusion of this theorem is that this RCT is informative about the average treatment response as long as a positive fraction of invited subjects participate in the experiment. The second key conclusion of this theorem is that unless the participation in the RCT is 100 percent among those invited, i.e. $P(P_i = 1|I_i = 1) = 1$, there is not point identification of the average treatment response. Note from the second corollary that even though there is not point identification of the ATE when $P(P_i = 1|I_i = 1) < 1$ there can be point identification of the sign of the ATE in many cases. Basically, the condition for point identification of the sign of the ATE is that the ATE in the subpopulation of subjects participating in the experiment is sufficiently large in magnitude relative to the fraction of subjects participating in the experiment to outweigh any possible ATE in the subpopulation of subjects not participating in the experiment. The width of the region of "ambiguity" as a function of $ATE_{exp}$ where the sign of the ATE is not identified is $2(M-m)P(P_i = 0|I_i = 1)$: the smaller this is, the more likely that the model will point identify the sign of the ATE.

**Remark 2.1** (Justifying the assumptions). *It is often the case that experimental studies report summary statistics that are used to suggest that randomization has "worked" because the observables of the subjects receiving each treatment (including perhaps the subjects in the control group when applicable) have similar distributional properties. Note that while this may bolster the case for assumption 2.2, it implies nothing about whether response is mean independent of participation in an RCT. A useful measure of whether response is mean independent of participation would be a comparison of the same distributional properties between participants and non-participants; depending on the nature of these covariates, this comparison may or may not be feasible.*[3] *However, it is important to note even this cannot "prove" that response is mean independent of participation because of the possibility of unobservables. In the conclusions we discuss this issue further.*

[3] For example, if the measurement of the covariates is invasive and non-standard then almost by definition of non-participation subjects who do not participate in the RCT will have missing data for these covariates. It is possible, however, to consider combining many different datasets.

2.1. **Selective participation in RCTs.** The reason that there is not point identification of the ATR and ATE in general is that the identified set accounts for the possibility that response is not mean independent of participation in the RCT. Indeed, the reasons for concern that response is not mean independent of realized treatment in observational data are basically exactly the same reasons that there should be concern that response is not mean independent of participation in an RCT. The simple reason is that participation in an RCT amounts to a gamble with the desire of selecting the experimental treatment.

This claim is consistent with a simple economic theory of how subjects decide whether to participate in an RCT. Suppose that subjects (or their agents; for example, their caregivers in a medical setting) have preferences over the treatment the subjects receive. Suppose in particular that the utility is the same as the actual outcomes that result from these treatments. This abstracts away from, for example, differences in the cost of treatments. It also requires that the subjects perfectly know the outcome that result from each of the treatments. Of course, the goal of the RCT is to learn the outcome that results from the treatments, so this assumption is almost certainly not literally true. But, it seems a good first approximation to motivate why the assumption that response is not mean independent of participation in an RCT is not necessarily credible.[4]

Suppose that treatments $\mathcal{D}_{ne} \subset \mathcal{D}$ are the non-experimental treatments available outside of the experiment, treatments $\mathcal{D}_e \subset \mathcal{D}$ are the experimental treatments available only in the experiment, and that treatments $\mathcal{D}_{rct} \subset \mathcal{D}$ are the treatments available in the experiment, which might include some non-experimental treatments. This abstracts away from the complications of being invited to participate by many different RCTs, which can

[4] This perfect foresight assumption can certainly be weakened at the expense of a more complicated (and realistic) model of decision making. This would be an interesting area of research to pursue.

happen in medical RCTs. Suppose for simplicity that subjects are risk-neutral expected utility maximizers, with utility equal to the outcome, and that they know the probability they will receive each treatment should they participate in the RCT. Then subject $i$ will participate if and only if $\sum_{d\in\mathcal{D}_{rct}} p_d y_i(d) \geq \max_{d\in\mathcal{D}_{ne}} y_i(d)$. A necessary condition is that $\max_{d\in\mathcal{D}_e} y_i(d) \geq \max_{d\in\mathcal{D}_{ne}} y_i(d)$, since otherwise participation in the RCT is dominated by not participating and being able to choose the optimal non-experimental treatment.[5] In particular, suppose that there is exactly one experimental treatment, $d_e$. Then a necessary condition for participation is that $y_i(d_e) \geq \max_{d\in\mathcal{D}_{ne}} y_i(d)$.

This suggests that participants in an RCT will tend to not be representative of the population of all subjects invited to participate, and indeed may tend to have greater treatment effect of the experimental treatment relative to other treatments than does the population of interest. But this does not necessarily imply anything about the response to the experimental treatment. It could be, for example, that participants in an RCT tend to have "bad" responses to all of the treatments, but just have a relatively better response to the experimental treatment than to the other treatments.

2.2. **Response is mean independent of participation: a necessary and sufficient condition for the usual RCT estimate.** If we maintain the assumption that response is mean independent of participation in the RCT, we have the following point identification result.

**Assumption 2.3.** *Assume that* $E(y_i(d)|P_i = 1, I_i = 1) = E(y_i(d)|I_i = 1) = E(y_i(d)|P_i = 0, I_i = 1)$ *for all treatments* $d \in \mathcal{D}$.

[5] Let $y_i^* = \max_{d\in\mathcal{D}_{ne}} y_i(d)$. Then re-write $\sum_{d\in\mathcal{D}_{rct}} p_d y_i(d) \geq \max_{d\in\mathcal{D}_{ne}} y_i(d)$ as $\sum_{d\in\mathcal{D}_e} p_d(y_i(d) - y_i^*) + \sum_{d\in\mathcal{D}_{rct}\cap\mathcal{D}_{ne}} p_d(y_i(d)-y_i^*) \geq 0$. The second sum is non-positive by definition of $y_i^*$, so $\sum_{d\in\mathcal{D}_e} p_d(y_i(d)-y_i^*) \geq 0$ and consequently $\max_{d\in\mathcal{D}_e} y_i(d) \geq y_i^*$.

**Theorem 2.2.** *Under assumptions 2.1, 2.2, and 2.3, $E(y_i(d))$ is point identified as $E(y_i|D_i = d, P_i = 1, I_i = 1)$.*

Note that if there is no heterogeneity in response then the mean independence part of assumption 2.3 holds. Note also that the resulting bound in theorem 2.2 is the same as the bound in theorem 2.1 when $P(P_i = 1|I_i = 1)$, so assuming that response is mean independent of participation can be interpreted as being equivalent to "assuming" that the participation rate among the invited is 100 percent. Also note that as the participation rate among the invited tends to 100 percent the bound in theorem 2.1 converges to the bound in theorem 2.2, implying that reporting the usual RCT estimate is a reasonably good approximation to the identified bound in theorem 2.1 when there is relatively high participation.

Moreover, suppose that the experimenter assumes that 2.1 and 2.2 hold, and then "assumes" that the average treatment response is point identified as $E(y_i(d)) = E(y_i|D_i = d, P_i = 1, I_i = 1)$. This is akin to assuming that the decision to participate in a trial among invitees is random, in the sense that response is mean independent of participation. Recall that under assumptions 2.1 and 2.2

$$\begin{aligned}
E(y_i(d)) &\underbrace{=}_{2.1} E(y_i(d)|I_i = 1) \\
&= E(y_i(d)|P_i = 1, I_i = 1)P(P_i = 1|I_i = 1) + E(y_i(d)|P_i = 0, I_i = 1)P(P_i = 0|I_i = 1) \\
&\underbrace{=}_{2.2} E(y_i|D_i = d, P_i = 1, I_i = 1)P(P_i = 1|I_i = 1) + E(y_i(d)|P_i = 0, I_i = 1)P(P_i = 0|I_i = 1).
\end{aligned}$$

Therefore, by algebra, for the condition $E(y_i(d)) = E(y_i|D_i = d, P_i = 1, I_i = 1)$ to hold it must be that either $P(P_i = 1|I_i = 1) = 1$, which seems non-generic, or $E(y_i(d)|P_i = 0, I_i = 1) = E(y_i|D_i = d, P_i = 1, I_i = 1) = E(y_i(d)|P_i = 1, I_i = 1)$. But this is precisely assumption 2.3. So under assumptions 2.1 and 2.2 the conventional estimate used in an RCT is equivalent to the assumption that response is mean independent of participation in the RCT.

More generally, it is sufficient for the conventional interpretation of an RCT as point identifying the average treatment response that three mean independence assumptions hold: response is mean independent of invitation, participation, and treatment assignment. An ideal RCT should satisfy the first and third assumptions but not necessarily the second assumption; the first and third assumptions are under the control of the experimenter but the second assumption is not.

2.3. **The average treatment response with an instrument for participation.** In this section it is shown that if there is a suitable instrument for participation it is possible to point identify the average treatment response. The instrument is a random variable $X_i$, which is discrete (for simplicity), and is defined through the following assumption. This variable is assumed observed.

**Assumption 2.4.** *Assume that $X_i$ is an observed instrument for all subjects who are invited to participate and that $E(y_i(d)|X_i = x, P_i = 1, I_i = 1) = E(y_i(d)|X_i = x, I_i = 1) = E(y_i(d)|X_i = x, P_i = 0, I_i = 1)$ for all treatments $d \in \mathcal{D}$ and all $x$ in the support of $X|I = 1$. Also assume that $P(P_i = 1|I_i = 1) > 0$. And assume also that $P(X_i = x|P_i = 1, I_i = 1) > 0$ and $P(X_i = x|P_i = 0, I_i = 1) > 0$ for all $x$ in the support of $X|I = 1$.*

In other words, response is mean independent of participation conditional on the instrument. We also assume that response is still mean independent of treatment assignment conditional now also on the instrument.

**Assumption 2.5.** *Assume that* $E(y_i(d)|X_i = x, P_i = 1, I_i = 1) = E(y_i(d)|D_i = d, X_i = x, P_i = 1, I_i = 1)$ *and further that* $P(D_i = d|X_i = x, P_i = 1, I_i = 1) > 0$ *for all treatments* $d \in \mathcal{D}$ *and all* $x$ *in the support of* $X|I = 1$.

Then we can write that

$$\begin{aligned} E(y_i(d)|P_i = 1, I_i = 1) &= \sum_x E(y_i(d)|X_i = x, P_i = 1, I_i = 1)P(X_i = x|P_i = 1, I_i = 1) \\ &= \sum_x E(y_i|D_i = d, X_i = x, P_i = 1, I_i = 1)P(X_i = x|P_i = 1, I_i = 1) \end{aligned}$$

and similarly

$$\begin{aligned} E(y_i(d)|P_i = 0, I_i = 1) &= \sum_x E(y_i(d)|X_i = x, P_i = 0, I_i = 1)P(X_i = x|P_i = 0, I_i = 1) \\ &= \sum_x E(y_i|D_i = d, X_i = x, P_i = 1, I_i = 1)P(X_i = x|P_i = 0, I_i = 1) \end{aligned}$$

where the second expression uses assumption 2.4. This establishes the following theorem.

**Theorem 2.3.** *Under assumptions 2.1, 2.4, and 2.5,* $E(y_i(d))$ *is point identified as* $\sum_x E(y_i|D_i = d, X_i = x, P_i = 1, I_i = 1)P(X_i = x|I_i = 1)$.

*Proof.* Simplify

$$\begin{aligned} &\left(\sum_x E(y_i|D_i = d, X_i = x, P_i = 1, I_i = 1)P(X_i = x|P_i = 1, I_i = 1)\right) P(P_i = 1|I_i = 1) \\ &+ \left(\sum_x E(y_i|D_i = d, X_i = x, P_i = 1, I_i = 1)P(X_i = x|P_i = 0, I_i = 1)\right) P(P_i = 0|I_i = 1) \end{aligned}$$

□

In other words, the average response to treatment $d$ is the weighted average of the average responses to treatment $d$ for subjects participating in the experiment and assigned treatment

$d$ who have covariates $x$, weighted by the probability that a subject invited to participate has covariates $x$. In the special case that $P(X_i = x|D_i = d, P_i = 1, I_i = 1) = P(X_i = x|I_i = 1)$ so that the distribution of the instrument is independent of participating in the experiment (and receiving treatment $d$) then this simplifies to the usual RCT result that $E(y_i(d)) = E(y_i|D_i = d, P_i = 1, I_i = 1)$ like in theorem 2.2. Otherwise the average treatment response needs to be re-weighted by the distribution of the instruments among all subjects invited to participate, not the distribution of the instruments among those actually participating.

2.4. **The average treatment effect with a "linear participation effect".** In this section it is shown that under a functional form assumption on the response function even if response is not mean independent of participation in the RCT it is possible to point identify the average treatment effect by "differencing out" the effect of participation. Suppose that $y_i(d) = y_{i0}(d) + \alpha_i$. This so far can always be done as we have not made any assumptions on $\alpha_i$. We will introduce an assumption implying that $\alpha_i$ captures all of the "endogeneity," or the selection effect of participation in the RCT, and so given these assumptions the linearity of $\alpha_i$ plays a crucial role. One can think of $y_0$ as the "regression" part and the $\alpha$ as the "error term" where the regression part is "exogenous" and all the selection is going through $\alpha$ which we can difference out since it does *not* depend on the treatment $d$. Then we can write that $y_i(d') - y_i(d) = y_{i0}(d') - y_{i0}(d)$, so the treatment effect does not depend on $\alpha_i$. The key identifying assumption is hence the following.

**Assumption 2.6.** *There exists $y_{i0}(d) : d \in \mathcal{D}$ and $\alpha_i$ with $y_i(d) = y_{i0}(d) + \alpha_i$ such that $E(y_{i0}(d)|P_i = 1, I_i = 1) = E(y_{i0}(d)|I_i = 1) = E(y_{i0}(d)|P_i = 0, I_i = 1)$ for all $d$.*

In other words, $\alpha_i$ captures all of the "endogeneity" of participation in the RCT. The restriction imposed by this assumption is that this endogeneity affects the response to all treatments equally. Then under the assumption that response is mean independent of treatment assignment, assumption 2.2, we have that $E(y_i|D_i = d, P_i = 1, I_i = 1) = E(y_i(d)|P_i = 1, I_i = 1) = E(y_{i0}(d)|P_i = 1, I_i = 1) + E(\alpha_i|P_i = 1, I_i = 1) = E(y_{i0}(d)|I_i = 1) + E(\alpha_i|P_i = 1, I_i = 1)$. Consequently, under the assumption that response is mean independent of invitation, assumption 2.1, $ATE(d', d) \equiv E(y_i(d') - y_i(d)) = E(y_i(d') - y_i(d)|I_i = 1)$ is point identified by $E(y_i|D_i = d', P_i = 1, I_i = 1) - E(y_i|D_i = d, P_i = 1, I_i = 1) = E(y_{i0}(d')|I_i = 1) - E(y_{i0}(d)|I_i = 1) = E(y_i(d') - y_i(d)|I_i = 1)$. This establishes the following theorem.

**Theorem 2.4.** *Under assumptions 2.1, 2.2, and 2.6, the average treatment effect* $ATE(d', d) \equiv E(y_i(d') - y_i(d))$ *is point identified as* $ATE(d', d) = E(y_i|D_i = d', P_i = 1, I_i = 1) - E(y_i|D_i = d, P_i = 1, I_i = 1)$. *However, the sharp identified set for* $E(y_i(d))$, *for* any one *treatment* $d$, *remains the same as in theorem 2.1.*

*Proof.* The identification of $ATE(d', d)$ follows from the previous discussion. The result that the sharp identified set for $E(y_i(d))$ remains the same as in theorem 2.1 is obtained by considering the response functions

$$y_{i0}(d) = \begin{cases} y_i & \text{if } D_i = d, P_i = 1, I_i = 1 \\ E(y_i|D_i = d, P_i = 1, I_i = 1) & \text{if } D_i \neq d, P_i = 1, I_i = 1 \\ E(y_i|D_i = d, P_i = 1, I_i = 1) & \text{if } P_i = 0, I_i = 1 \\ E(y_i(d)|I_i = 1) & \text{if } I_i = 0. \end{cases}$$

and

$$\alpha_i = \begin{cases} 0 & \text{if } D_i = d, P_i = 1, I_i = 1 \\ 0 & \text{if } D_i \neq d, P_i = 1, I_i = 1 \\ [m, M] - E(y_i | D_i = d, P_i = 1, I_i = 1) & \text{if } P_i = 0, I_i = 1 \\ 0 & \text{if } I_i = 0. \end{cases}$$

These response functions $y_{i0}(\cdot)$ and $\alpha_i$ add up to the same response function used to establish sharpness in the proof of theorem 2.1. Therefore they are consistent with the data, and assumptions 2.1 and 2.2, and achieve any point in the identified set. They also satisfy assumption 2.6.[6] □

Note that this provides a justification for running an RCT with a control group even when the average response in the population of interest to the treatment given to the control group is already known from previous experiments, since the average response in the subpopulation of subjects that participate in the RCT to the treatment given to the control group is likely not known. It is useful to note that this does not require a control group that receives only a placebo, and indeed the medical ethics of placebo control groups is in question (i.e., Rothman and Michels (1994)). If the average treatment effect of treatment 1 versus a placebo is already known, it is enough by the linearity of expectations to point identify the average treatment effect of treatment 2 versus treatment 1 in order to point identify the average treatment effect of treatment 2 versus a placebo. The approach in this section can be used on an RCT

[6] Note that the restriction imposed by assumption 2.6 is that $\alpha_i$ is the same across treatments, so there is a restriction in the identified sets for $E(y_i(d))$ across different $d$; intuitively if the $\alpha_i$ required to rationalize the data for some point in the identified set for $E(y_i(d))$ for some $d$ is one unit bigger it means that the response to all treatments $d'$ is one unit bigger among the subpopulation with $P_i = 0, I_i = 1$. So as the previous discussion shows this is a sufficient restriction to point identify the ATE.

involving only treatments 1 and 2, and then by this argument the average treatment effect of treatment 2 versus a placebo can be inferred logically.

Note that this result places a strong functional form assumption on the response function. Unless the participation effect enters the response function linearly, differencing does not remove the participation effect. In many cases a non-linear effect of participation on response is plausible. For example, in an economics RCT it could be that the participation effect captures the "motivation" of the subject and that the response to some treatments depends on motivation while the response to other treatments does not depend on motivation. Similarly, in a medical RCT it could be that participation captures the baseline health status (e.g., "performance status" in the case of chemotherapy) of the subject and that some treatments like chemotherapy provide good outcomes for subjects who come into the experiment already reasonably healthy, but provide bad outcomes for subjects who come into the experiment less healthy. So as with any assumption, it is important to consider whether the assumption of a linear participation effect is credible in any given RCT.

## 3. Model 2: an RCT with unobserved invitation

The analysis in the previous section has been of a model in which the participation rate and invitation rate is observed. This is a reasonable modeling assumption for some RCTs, but is not a reasonable modeling assumption for other RCTs. For example, consider an RCT that is run at a job training center. In this setting the experimenter observes which people elect to participate conditional on being invited to participate in an RCT, but does not observe the fraction of the population of interest that never show up to the job training center. This is because the experimenter does not observe anything about the people who do not come to the job training center. It could be, for example, that the people who come to the job training center are an arbitrarily small fraction of the population of interest. In order to model this we introduce another decision that occurs before any of the three decisions of the previous model. This is the decision of the subject to attend whatever "meeting" is necessary to get an invitation to participate in the RCT. The variable indicating attendance is $A_i$. This may partially capture things like the motivation, ability, or other baseline characteristics of the subject that are correlated with the response. For example, this is the decision of the subject to go to a job training center. The previous model is a special case of this model in which $A_i = 1$ for all subjects $i$. This is reasonable, for example, in a medical setting where patients must receive some level of medical care for their disease and so all patients have $A_i = 1$, and also in certain types of economics RCTs, especially those where the experimenter actively invites participation directly from the population of interest.

Again using the law of iterated expectation we get that

$$E(y_i(d)) = \underbrace{E(y_i(d)|A_i = 1)}_{(1)}\underbrace{P(A_i = 1)}_{(2)} + \underbrace{E(y_i(d)|A_i = 0)}_{(3)}\underbrace{P(A_i = 0)}_{=1-(2)}.$$

So, as before, the average response to treatment $d$ can be decomposed as the weighted average of the average treatment responses for subjects attending, (1), and not attending, (3), the meeting where the weight is the probability of meeting, (2). It is not credible to assume that response is mean independent of attendance for the same reasons that it is not credible to assume that response is mean independent of participating conditional on participating. The difference now is that the experimenter does not even observe subjects with $A_i = 0$, so does not know $P(A_i = 1)$. Therefore, we have the result that this type of RCT is (basically) completely uninformative about the average treatment response. This is because $P(A_i = 1)$ could be arbitrarily small, and the data reveals nothing about $E(y_i(d)|A_i = 0)$.

**Theorem 3.1.** *In this model the (closure of the) identified set for $E(y_i(d))$ is that $E(y_i(d)) \in [m, M]$.*

On the other hand, if for some reason it is credible to assume that $E(y_i(d)|A_i = 0) = E(y_i(d)|A_i = 1)$ then informative identification can proceed by using the model of the previous section on $E(y_i(d)|A_i = 1)$ and extrapolating to $E(y_i(d)|A_i = 0)$ using this assumption that response is mean independent of attendance. For more about this situation in which the population of interest is not observed see Manski (1996).

## 4. Comparing RCTs to observational data

This section compares RCTs to observational data. Obviously since the RCT of model 2 is (basically) completely uninformative about the response functions, observational data cannot be worse than an RCT of model 2. Therefore this section focuses on the RCT of model 1. Further, we consider the identification of the model with data from an RCT as derived in theorem 2.1 since with the additional assumptions we entertain the RCT point identifies the objects of interest so necessarily is better than observational data.

**Theorem 4.1.** *Under no assumptions, with observational data the identified set for $E(y_i(d))$ is $E(y_i(d)) \in E(y_i|D_i = d)P(D_i = d) + [m, M]P(D_i \neq d)$.*

*Further, the identified set for $ATE(d', d)$ is $ATE(d', d) \in E(y_i|D_i = d')P(D_i = d') - E(y_i|D_i = d)P(D_i = d) + [mP(D_i \neq d') - MP(D_i \neq d), MP(D_i \neq d') - mP(D_i \neq d)]$.*

These are the same bounds as derived in the general case of observational data; see Manski (2007).

4.1. **Identifying the sign of the ATE.** By algebra the lower bound on the identified set for $ATE(d', d)$ with observational data can be written as $(E(y_i|D_i = d') - M)P(D_i = d') + (m - E(y_i|D_i = d))P(D_i = d) + (m - M)(1 - P(D_i = d') - P(D_i = d))$ and the upper bound can be written as $(E(y_i|D_i = d') - m)P(D_i = d') + (M - E(y_i|D_i = d))P(D_i = d) + (M - m)(1 - P(D_i = d') - P(D_i = d))$. Assuming that $m < M$, which is generic since otherwise there is no uncertainty about outcomes, these expressions make it clear that the sign of $ATE(d', d)$ is not point identified with observational data except possibly in the special case that $1 - P(D_i = d') - P(D_i = d) = 0$. Otherwise, for the lower bound the first two terms are non-positive and the last term is negative, and similarly for the upper bound the first two terms are non-negative and the last term is positive. Therefore both strictly positive and strictly negative values for ATE are in the identified set. In case $1 - P(D_i = d') - P(D_i = d) = 0$ and $P(D_i = d) > 0$ and $P(D_i = d') > 0$ then a non-negative ATE is point identified exactly in case $E(y_i|D_i = d') = M$ and $E(y_i|D_i = d) = m$. A non-positive ATE is point identified exactly in case $E(y_i|D_i = d') = m$ and $E(y_i|D_i = d) = M$. Note in particular that it is never the case that a zero ATE can be ruled out with observational data, since it is always consistent with the data that $y_i(d') = y_i = y_i(d)$ for all $d', d$.

On the other hand, recall that under reasonable conditions the sign of the ATE is point identified with data from an RCT. Therefore under the assumptions of theorem 2.1, the RCT is valuable above observational data for the purposes of identifying the sign of the ATE.

4.2. **The width of the identified sets.** Note also that the width of the identified set for the average treatment response $E(y_i(d))$ with RCT data is $(M-m)P(P_i=0|I_i=1)$ while with observational data it is $(M-m)P(D_i \neq d)$. The width depends on the treatment considered with observational data but does not with RCT data. This makes it difficult to give a general comparison of identification with RCT data and observational data in terms of average treatment response, because the comparison depends on the treatment considered and the details of how treatments are selected in the observational data.

So, in order to compare the two types of data in an apparently general way we consider the sum of the widths of identified sets across all treatments, which is a measure of the total "uncertainty" that remains about the average treatment responses. With RCT data this is $(M-m)|\mathcal{D}|P(P_i=0|I_i=1)$ while with observational data this is $\sum_{d\in\mathcal{D}}(M-m)(1-P(D_i = d)) = (M-m)|\mathcal{D}|(1-\frac{1}{|\mathcal{D}|})$. Therefore the RCT is as good as observational data if and only if $P(P_i=0|I_i=1) \leq 1-\frac{1}{|\mathcal{D}|}$ or equivalently $|\mathcal{D}|P(P_i=1|I_i=1) \geq 1$. That is, the RCT is preferred when there are many treatments and/or when there is high participation in the RCT among those invited.

A similar result obtains for the sum of the widths of the identified sets for the average treatment effects. The identified sets for $ATE(d',d)$ is the difference in the identified sets for $E(y_i(d'))$ and $E(y_i(d))$ under any of the assumptions entertained here since there are no restrictions across treatments. Therefore the width of the identified set for $ATE(d',d)$ is the sum of the widths of the identified sets for $E(y_i(d'))$ and $E(y_i(d))$. Let $\mathcal{H}(\cdot)$ be the identified set for its argument. Therefore we have that $\sum_{d\in\mathcal{D}}\sum_{d'>d} width(\mathcal{H}(ATE(d',d))) = \sum_{d\in\mathcal{D}}\sum_{d'>d} width(\mathcal{H}(E(y_i(d')))) + width(\mathcal{H}(E(y_i(d)))) = (|\mathcal{D}|-1)\sum_{d\in\mathcal{D}} width(\mathcal{H}(E(y_i(d)))$. Therefore the comparison on the basis of the sums of the widths of the identified sets for the average treatment effects is the same as on the basis of the sums of the widths of the identified sets for the average treatment responses.

**Remark 4.1** (Combining RCT and observational data). *We have derived above various bounds on treatment response using RCT data. If one has access to observational data, then bounds can be combined. A simple way to do that is to obtain bounds on, for example, the ATR using both RCT and observational data and then form the intersection of these bounds to get the overall bound on ATR. This would be a simple and effective approach to combining both data sources, although the resulting intersection of bounds may not be sharp. The sharp bounds will in general depend on potentially subtle issues relating to how the RCT and observational populations "overlap" in their relationship to the population of interest; deriving these bounds is left to future work.*

## 5. Empirical illustration

In this section we illustrate our identification results using a recent analysis of recruitment into medical RCTs in GMHK. As far as we know there is no similar analysis of economics RCTs, perhaps because of the relatively young age of experiments in economics. Nevertheless we suppose, as we have done throughout the paper, that at least as a stylized fact the conduct of experiments in medicine and in economics is similar, and so our results for medical RCTs are informative also for economics RCTs.

GMHK study 172 medical RCTs published over the course of a year in four major medical journals. In these RCTs the median eligibility fraction, the fraction of the potential participants who are eligible to enroll in the study after screening (roughly analogous to invitation in our model), is 65%. The interquartile range is 41-82. These figures are based on the 48 studies that report the necessary data in the publication. The median enrollment fraction, the fraction of eligible participants who actually enroll (roughly analogous to participation among the invited in our model), is 93%. The interquartile range is 79-100. These figures are based on the 74 studies that report the necessary data.[7] They note that 20 studies report an enrollment fraction of 100%, which they suggest is implausible and may be due to studies which conflate eligibility and participation in their reporting. The median recruitment fraction, the product of these two fractions, is 54%. The interquartile range is 32-77. These figures are based on 81 studies that report the necessary data.

We suppose in our empirical illustration that the outcome is binary. In an economics RCT this might be school attendance or employment, and in a medical RCT this might be survival at 6 months. Consequently we take $m = 0$ and $M = 1$ in our model above.[8] In table 1 we consider how the participation rate among those invited, $P(P_i = 1|I_i = 1)$, relates to the identification of the sign of the average treatment effect. We suppose that there are two treatments of interest, and that some RCT satisfying the conditions of theorem 2.1 reveals the experimental ATE, $ATE_{exp}(d', d') \equiv (E(y_i|D_i = d', P_i = 1, I_i = 1) - E(y_i|D_i = d, P_i = 1, I_i = 1))$. Recall this is the ATE on the subpopulation that actually participates in the experiment. We study the smallest experimental ATE such that according to corollary 2.2 the ATE in

[7] Note that this implies less than one-half of studies report this data; it is not obvious which direction the resulting median is biased from the median in the population. Ironically, this is itself because of a partial identification issue. It could be that studies with a low enrollment fraction are more likely to report that, because it may threaten the validity of the study and so it is worth reporting. Alternatively, it could be that studies with a low enrollment fraction are less likely to get published in a major medical journal, exactly because a low enrollment fraction may threaten validity.

[8] Note that the additional assumption of a discrete outcome rather than a continuous outcome does not affect the sharpness of the bounds. In proving sharpness we exhibited response functions that are the same for all people in certain subpopulations (e.g., the response to treatment $d$ is the same for everybody with $I_i = 0$). These exhibited response functions in general will not take values compatible with discreteness of the outcome, but necessarily the response functions take values in the convex hull of the set of outcomes. It is trivial to simply partition any given subpopulation further and assign people in that sub-subpopulation to have outcomes compatible with the discreteness of the outcome, and such that on average that subpopulation has the same outcome as does the subpopulation in our sharpness proofs.

the population is point identified to be non-negative. If the participation rate is too low it is never possible to point identify the population ATE to be non-negative, and we indicate this in the table by "n.p." for not possible.

Each row of the table provides a possible participation rate among the invited and the corresponding smallest experimental ATE that point identifies the population ATE to be non-negative. So if the participation rate among invited is 60%, for example, then the experimental ATE must be at least as great as $\frac{2}{3}$ in order to point identify the population ATE to be non-negative. If the participation rate is strictly less than 50% then it is not possible to point identify the population ATE to be non-negative, because even if the experimental ATE were 1, the largest possible, the ATE in the subpopulation that does not participate but is invited could be $-1$, which would result in a negative population ATE. Note that the marginal gain in identifying power is greatest when the participation rate among the invited is low, in the sense that the derivative of the smallest $ATE_{exp}$ implying the population ATE is non-negative is (when it exists) $-P(P_i = 1|I_i = 1)^{-2}$, and so is decreasing in magnitude in the participation rate among the invited. Consequently, there is relatively less gain from increasing participation among the invited from 90% to 100% and relatively more gain from increasing participation from 50% to 60%.

Recall that the median recruitment fraction is 54%, and that the 25th quantile of the recruitment fraction is 32%. Although the recruitment fraction accounts for both eligibility and enrollment, recall from the discussion above in the context of RCTs of heart failure treatments that many RCTs have selective eligibility standards, and so the assumption that response is mean independent of invitation (or, analogously, eligibility in the context of GMHK) may not be appropriate for all RCTs. In those cases, as discussed before, it is

| Participation rate among invited | Smallest $ATE_{exp}$ implying $ATE \geq 0$ |
|---|---|
| 10% | n.p. |
| 20% | n.p. |
| 30% | n.p. |
| 40% | n.p. |
| 50% | 1 |
| 60% | $\frac{2}{3} \approx .67$ |
| 70% | $\frac{3}{7} \approx .43$ |
| 80% | $\frac{1}{4} = .25$ |
| 90% | $\frac{1}{9} \approx .11$ |
| 100% | 0 |

Table 1. Effect of participation rate on identification of the sign of the average treatment effect; n.p. = not possible, there is no $ATE_{exp}$ implying $ATE \geq 0$.

useful to re-define invitation and participation in our model so that all selection happens exclusively through participation. Moreover, only roughly one-half of the RCTs studied by GMHK report the recruitment fraction. Overall, these concerns point to the possibility, translated to our model, that many medical RCTs effectively have a participation rate that is less than 50%. And in those cases the results in table 1 show that it is not possible to point identify the sign of the average treatment effect for a binary outcome, no matter what the experimental data reveals.

This means that as a practical matter it is important to think carefully about the assumptions maintained about participation in RCTs. Per the discussion above in section 2.2 the usual RCT estimate is equivalent to "assuming" that the participation rate among the invited is 100 percent, and so implicitly entails stronger assumptions when the participation rate among the invited is substantially less than 100 percent as in many of the RCTs studied by GMHK. Of course it is possible that response is mean independent of participation, an assumption that depends on the details of the RCT, in which case the usual RCT estimate is credible. When that assumption is not credible the usual RCT estimate fails to account for the selectivity of participation, and in many RCTs this effect is potentially strong enough to imply that not only is the ATE not point identified, but even the sign of the ATE is not point identified. We suggest that RCTs should either report the sorts of bounds from theorem 2.1 in addition to point estimates, or use our alternative identification strategies based on an instrument for participation or a functional form assumption of a linear participation effect. If this is not possible it is useful to at least provide evidence for why the usual RCT estimate is justified in terms of the selectivity of participation. We discuss this further in the conclusions.

## 6. Conclusions

This paper studies the question of what we can learn about the average treatment response (ATR) and average treatment effect (ATE) with data from a randomized controlled trial under weak assumptions, and compares the results to what we can learn about these same objects with observational data. We focus on the problem of selection into an RCT, which happens because subjects are not randomly assigned from the population of interest to participate in the RCT. This is similar to the usual selection problem that treatments are not randomly assigned in observational data.

The key difference between the selection into an RCT problem and the usual selection problem is that, once a subject does participate in the RCT, it is randomly assigned a treatment. So there is at least some subpopulation for whom we can assume that response is mean independent of treatment. This is not true with observational data. On the other hand observational data has the potential advantage of providing information about all subjects in the population of interest, albeit subject to the selection problem. We show that this tradeoff for a given treatment may favor either the RCT or observational data in terms of the width of the identified set for the ATR. The intuition for this result is that while RCT data is stipulated to have high internal validity on the subpopulation of interest, when there is *heterogeneity in the ATR*, it could be that the subjects who do not participate in the RCT have different ATR. It is possible that even though observational data has the usual selection problem it has data on a greater fraction of the population of interest so that it has narrower identified sets.

We have also provided three conditions under which RCT data does point identify the object of interest. These conditions are that response is mean independent of participation, or that there are instruments for participation, or that there is a linear effect of participation on response. Depending on the particular data one has, other sets of assumptions can be used such as monotonicity of response functions, or other assumptions common in the partial identification literature.

Of course, the key population quantities that determine whether to prefer RCT or observational data are the probability that a subject that is invited to participate in an RCT actually does participate and the difference in the mean response between those who participate and those who do not participate. In case all invited subjects do participate, there is point identification and hence RCT data allows to fully learn the treatment effects. Similarly, if participation is sufficiently random that response is mean independent of participation, so assumption 2.3 holds, there is also point identification. These two conditions depend on the details of a particular RCT so no generic statements can be given here. However, there is suggestive evidence that in many RCTs there is low participation and those who participate are different from those who do not participate. Rothwell (2005) suggests that as few as 0.001 percent of the population of interest may participate in a certain type of medical RCT. This figure accounts for subjects who do not participate, for example, because they are (roughly, translating to our model) not "invited" to participate, so this may understate $P(P_i = 1|I_i = 1)$. Nevertheless this suggests considerable non-participation of subjects.[9]

Low participation rates are also characteristic of some RCTs in economics. For example, Banerjee and Duflo (2009) report an experiment of a "no legal strings attached" gift of "between $25 and $100" as part of the Bandhan microfinance program in India. Approximately 20% of the invited subjects (translating roughly to our model) rejected the gift. Another example is the experimental study of the Job Training Participation Act (JTPA) conducted by the Manpower Demonstration Research Corporation. In this experiment, Doolittle and Traeger (1990) and Heckman (1992) report that more than 90% of invited training centers refused to participate.

[9] Recall also that we assume that response is mean independent of invitation; the analysis in Rothwell (2005) suggests this may not always be a credible assumption. Consequently the 0.001 figure may be closer to the "true" non-participation after accounting for the fact that invitation is correlated with response.

A low participation rate does not necessarily imply, however, that response is not mean independent of participation, although perhaps it does raise questions. In many cases by definition of not participating in the RCT limited data is available on subjects who do not participate. Nevertheless, it may be possible to compare the characteristics of subjects who participate with, for example, population data from other sources, in order to get a sense of whether participation seems likely to be related to response. This sort of analysis is conducted by Steg, Lopez-Sendon, Lopez de Sa, Goodman, Gore, Anderson Jr, Himbert, Allegrone, and Van de Werf (2007) in a meta-analysis of RCTs of treatments for acute myocardial infarction ("heart attack"). They find that the characteristics of patients who are eligible for an RCT but do not participate are "worse" than of patients who actually participate. The same pattern holds for the observed outcomes.[10] Rothwell (2005) provides data that suggests that even this comparison may not be enough because of characteristics that are unobserved in the data but are related to participation. He reports that in a RCT of endarterectomy to prevent stroke that roughly 3% of the patients were randomized into receiving the endarterectomy but "did not have surgery because their surgeon and/or anesthetist judged them to be too frail." This group had a distribution of observables similar to that for the rest of the participants in the RCT but did have a much higher subsequent rate of stroke compared to the patients participating in the RCT but not receiving the endarterectomy. This suggests that showing that observable characteristics of participants in an RCT are similar to those not participating may not be enough to establish that response is mean independent of participation.

[10] Their analysis also suggests that invitation does not satisfy the response mean independent of invitation assumption.

A similar conclusion that response is not mean independent of participation can be drawn about the experimental study of the JTPA using the data reported by Doolittle and Traeger (1990, Tables 5.4-5.5). In that study the training centers that participated tended to, among other things, be geographically un-representative of all centers, serve a smaller number of terminees from Title IIA of the JPTA, and had greater adult employment rates.

There is further suggestive evidence of a selection into an RCT problem provided by experiments which have a non-random procedure for assigning subjects to treatments. Of course, this violates in general assumption 2.2, but provides evidence about how and why subjects come to participate in an RCT. An example of such a procedure is to assign treatment according to whether the day admitted is even or odd. Pocock (1983) reports some medical experiments with such a procedure where there is a considerable imbalance in the number of patients receiving each treatment, suggesting that patients (or their agents) had preferences over which treatment to receive and manipulated their participation in the RCT accordingly. It stands to reason the same motivations would result in selection into an RCT.

Note that these examples depend on covariates, and if there are observed covariates that suitably explain participation then the instrumental variable strategy of section 2.3 can be used.

The analysis of this paper suggests that in reporting the results of an RCT it is useful to consider reporting the bounds on the ATR and ATE as derived in this paper. If this is not possible then it is useful to report information to the extent possible on the invitation rate; how subjects are invited to participate; the characteristics of those who are invited, and are not invited; the participation rate of those invited; and the characteristics of those who participate, and are invited but do not participate. This view is partly seen in the CONSORT statement (i.e., Moher, Hopewell, Schulz, Montori, Gotzsche, Devereaux, Elbourne, Egger, and Altman (2010) and Schulz, Altman, and Moher (2010)), a major set of guidelines for medical RCTs, which states that "[a] comprehensive description of the eligibility criteria used to select the trial participants is needed to help readers interpret the study" and "[a] description of the method of recruitment, such as by referral or self selection (for example, through advertisements), is also important in this context."

It is also worth noting that, for the purposes of identification, it is the invitation and participation *rates* that matters, not the absolute number of participants. This suggests that the emphasis in designing RCTs should be on high participation rates, not simply a large number of participants. This is because participants in a large RCT can be equally or less representative of the population of interest than participants in a small RCT if the two RCTs differ in their emphasis on recruitment of a representative subpopulation. Of course, statistical precision is improved with a larger sample. But a statistically precise estimate of a less informative identified set may be less preferred than an imprecise estimate of a more informative identified set.

# References

Banerjee, A. V., and E. Duflo (2009): "The experimental approach to development economics," *Annual Review of Economics*, 1(1), 151–178.

Doolittle, F., and L. Traeger (1990): "Implementing the National JTPA Study," .

Gross, C., R. Mallory, A. Heiat, and H. Krumholz (2002): "Reporting the recruitment process in clinical trials: who are these patients and how did they get there?," *Annals of Internal Medicine*, 137(1), 10–16.

Heckman, J. J. (1992): "Randomization and social policy evaluation," in *Evaluating Welfare and Training Programs*, ed. by C. F. Manski, and I. Garfinkel, chap. 5. Harvard University Press.

——— (1996): "Randomization as an instrumental variable," *The Review of Economics and Statistics*, 78(2), 336–341.

Heiat, A., C. Gross, and H. Krumholz (2002): "Representation of the elderly, women, and minorities in heart failure clinical trials," *Archives of Internal Medicine*, 162(15), 1682–1688.

Manski, C. F. (1996): "Learning about treatment effects from experiments with random assignment of treatments," *The Journal of Human Resources*, 31(4), 709–733.

——— (2007): *Identification for prediction and decision*. Harvard University Press.

Moher, D., S. Hopewell, K. Schulz, V. Montori, P. Gotzsche, P. Devereaux, D. Elbourne, M. Egger, and D. Altman (2010): "CONSORT 2010 Explanation and Elaboration: updated guidelines for reporting parallel group randomised trials," *British Medical Journal*, 340, c869.

Pocock, S. J. (1983): *Clinical trials: A practical approach*. John Wiley & Sons.

Rothman, K., and K. Michels (1994): "The continuing unethical use of placebo controls," *New England Journal of Medicine*, 331(6), 394–398.

Rothwell, P. (2005): "External validity of randomised controlled trials: To whom do the results of this trial apply?," *The Lancet*, 365(9453), 82–93.

Schulz, K., D. Altman, and D. Moher (2010): "CONSORT 2010 Statement: updated guidelines for reporting parallel group randomised trials," *British Medical Journal*, 340, c332.

Steg, P., J. Lopez-Sendon, E. Lopez de Sa, S. Goodman, J. Gore, F. Anderson Jr, D. Himbert, J. Allegrone, and F. Van de Werf (2007): "External validity of clinical trials in acute myocardial infarction," *Archives of Internal Medicine*, 167(1), 68–73.

Van Spall, H., A. Toren, A. Kiss, and R. Fowler (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(11), 1233–1240.

Weijer, C., S. Shapiro, K. Glass, and M. Enkin (2000): "For and against: Clinical equipoise and not the uncertainty principle is the moral underpinning of the randomised controlled trial," *British Medical Journal*, 321, 756–758.