Risk and evidence of bias in randomized controlled trials

Alex Eble and Peter Boone*

June 2012

PRELIMINARY DRAFT - NOT FOR CIRCULATION; COMMENTS WELCOME

ABSTRACT:

The randomized controlled trial (RCT) has been a heavily utilized research tool in medicine for over 60 years. It has enjoyed recent popularity in the social sciences, where it has been used to evaluate questions of both policy and theory. The early economics literature on RCTs invokes the medical literature, but seems to ignore a large body of this literature which studies the past mistakes of medical trialists and links poor trial design, conduct and reporting to exaggerated estimates of treatment effects. Using a few consensus documents on these issues from the medical literature, we design a tool to evaluate adequacy of reporting and risk of bias in RCT reports. We then use this tool to evaluate all reports of RCTs published in a set of 50 major economics journals between 2000 and 2009 alongside a sample of reports of RCTs published in medical journals over the same time period. We find that economics RCTs fall far short of the recommendations for reporting and conduct put forth in the medical literature, while medical trials stick fairly close to them, suggesting risk of exaggerated treatment effects in the economics literature.

* Boone: Centre for Economic Performance, London School of Economics and Effective Intervention. Eble: Brown University and Effective Intervention. The authors would like to thank Simon Johnson and especially Diana Elbourne and Miranda Mugford for helpful conversations, and participants at the Centre for Economic Performance 2009 Annual Conference and the Georgetown University Quantitative Models Group. Eble gratefully acknowledges the financial support of the US National Science Foundation as well as helpful comments from Samuel Brown, Morgan Hardy, Vernon Henderson, Paul Musgrave, Gareth Olds, Anja Sautmann, Tim Squires, David Weil, Hyunjoo Yang, and participants at the Centre for Economic Performance's 2009 annual conference, Brown University micro lunch, and Georgetown University Quantitative Models Group. All remaining errors are our own.

Section I: Introduction

The pedigree of the scientific experiment goes back millennia. In Daniel 1:8-16, King Nebuchadnezzar presides over what is arguably the first trial on record. The King orders a group of his subjects to eat rich meat and drink wine while another group is made to adhere to a vegetarian diet. After a ten day period, outcomes were compared in order to assess the relative merit of each "treatment" and the vegetarians won out. Many approximations of the randomized controlled trial (RCT) have been conducted since, (Twyman 2004) but the dawn of the current era of the RCT is largely associated with the publication of a set of articles by the British Medical Research Council testing the efficacy of pertussis vaccines on pulmonary tuberculosis in the 1940's. (Bell 1948) Since then, the field has blossomed. The method has been shown by several studies to yield more reliable treatment effect estimates than observational studies, and as a result the RCT has been adopted in several scientific fields as the "gold standard" of evidence quality and relied upon by government agencies in determining suitability of drug treatments. (Vader 1998) Nonetheless, decades of use and scrutiny have revealed numerous potential problems in the execution and review of RCTs centering on a set of six concerns related to trial conduct and analysis. The problems central to each of the six concerns have been associated with exaggerated treatment effects relative to studies whose design anticipates and attempts to prevent such problems. Stemming from these findings, a few consensus documents have been developed to provide give guidance on how best to design, conduct and report trials in order to minimize the risk of such problems biasing the results.

In the past decade, the RCT has been widely adopted by economists - largely on the virtue of its "clean" identification of causal relationships - and has been used to evaluate literally hundreds of questions of both academic and policy interest. The Abdul Latif Jameel Poverty Action Lab (JPAL), a major center for the conduct of RCTs in economics, had either finished or was currently conducting over 200 RCTs by the beginning of 2010. (Parker 2010) Other groups at Yale, Berkeley and elsewhere have

recently formed with similar missions. Though economists mention and often cite the medical literature as the inspiration for this approach, (Banerjee 2007) surprisingly few published reports of economics trials published before 2010 reference any of the wealth of medical articles on the pitfalls which have been shown to lead to biased results or on the means by which to reduce such biases.

Our research question is this: have trialists in economics taken the necessary steps to avoid the bias-inducing pitfalls that the medical literature has identified? Below, we briefly summarize the medical literature on bias in RCTs. We have used this literature to develop an instrument (henceforth, the "grid") with which to evaluate adequacy of reporting and risk of the six aforementioned biases (henceforth referred to simply as "bias") in published accounts of RCTs. Though we recognize it is an open question to what extent the standards from medicine can be meaningfully applied to economics, we argue that the medical standards offer a very clear link between certain RCT conduct decisions and treatment effect exaggeration, and that there is no reason not to use this information. After the discussion, we then use the grid to evaluate a set of journal articles documenting the results of RCTs in both economics and medicine.[1] We find that many of the economics articles provide insufficient information for the reader to assess the quality of the evidence presented and several others fall into the same traps that have previously skewed the results of trials in medicine. We finish by suggesting a similar set of guidelines for trialists in economics to follow when conducting and evaluating RCTs and offering a few paths for future research.

The rest of the paper proceeds as follows: section II discusses trials in medicine, the problems in design and implementation that medical trialists have come up against, the state of the art in medical RCTs in terms of avoiding these problems, and the advent of trials in economics. Section III describes the

[1] Economists have long been concerned with the same issues of identifying causality that led medical scientists to use RCTs, and there is a rich history of economists conducting experiments, both in the laboratory and beyond. Our analysis focuses exclusively on the use of prospectively designed, relatively large-scale RCTs in economics which have been in vogue only for the last decade and whose mission, arguably, mirrors the "Phase III" trial in medicine. They differ from previous experiments in economics in both their scale and mission.

methodology of our review, including how the grid was crafted, how eligibility was determined, how the search for papers was conducted, and how the analysis was implemented. Section IV presents our results, Section V suggests a way forward, and Section VI concludes.

Section II: Trials in medicine and economics

The history of randomized trials is well documented elsewhere (Collier 2009; Meldrum 2000), so we provide only a brief discussion of their development to motivate our analysis. Though the first parallel group study ostensibly dates to pre-Christian times as discussed in Section I, trials have only been broadly consumed by the medical community since the 1940's. As early as 1980 the RCT was recognized for its superior identification of causal relationships relative to other research designs, (Vader 1998) confirmed empirically in a series of meta-analyses which showed that nonrandomized studies yielded larger effect sizes than those found in randomized trials.(Ioannidis et al. 2001)

Subsequent analysis of the evidence provided by RCTs revealed that errors in design or analysis could lead to exaggerated treatment effect estimates in trials themselves. A series of studies investigated the relationship between methodological quality of RCTs and measured effect size, beginning with a landmark 1995 article published in the Journal of the American Medical Association. In this article, Schulz and colleagues found that trials with inadequately concealed treatment allocation estimated up to 30% larger treatment effects than well designed studies with adequate allocation concealment. (K. F. Schulz et al. 1995) This finding and others similar to it instigated a larger movement to improve and standardize both methods of reporting RCTs and methods of scrutinizing them.

In the 1990's, two groups began independently working on establishing a set of reporting standards to be used in publication of randomized trials, the goal of which was to ensure that readers of articles reporting the results of RCTs had sufficient information to confirm or refute that the trial had in fact been carried out in a manner that would lead to unbiased results. Their combined efforts resulted in

the CONSORT Statement, a set of guidelines for publication of reports of randomized controlled trials. Adherence to these standards is now required by most editors of major medical journals. (K. Schulz et al. 2010)

The Cochrane Collaboration, another arm of this movement, is an international organization which facilitates systematic review and meta-analysis of published studies in order to draw overall conclusions about efficacy of various treatments. It publishes a handbook that instructs authors how to conduct these meta-analyses which includes a section on how to evaluate the quality of evidence provided by RCTs. The handbook, which is updated frequently, has been used to conduct over 6,200 systematic reviews of trials, assessing the quality of evidence in hundreds of thousands of scholarly articles. (The Cochrane Collaboration 2010)

The Cochrane handbook and CONSORT guidelines together offer a thorough discussion of the biases and other problems identified as historically posing a significant problem to obtaining accurate results from RCTs. These biases include selection, performance, detection, attrition, reporting and sample size biases. (Jüni et al. 1999; Higgins, Green, and Cochrane Collaboration 2008; Moher et al. 2010) The remit of each of these issues, of course, is vast, and thorough exploration of any of them is beyond the scope of this article. Instead, we discuss each issue briefly and cite a few major studies which demonstrate the implications of study design which fails to address these potential pitfalls.

Selection bias

Selection bias is concerned with two main issues: one, that systematic differences arise between the sampling population and the sample drawn and, two, that systematic differences arise between treatment groups at the outset of the trial, usually due to individuals either tampering with or predicting the allocation sequence. A review of several meta-analyses which aggregated the results of Schulz and others found that “odds ratios were approximately 12 percent more positive in trials without adequate

allocation sequence generation" and that "trials without adequate allocation concealment were approximately 21 percent more positive than trials with adequate allocation concealment". (Gluud 2006)

CONSORT asserts that "authors should provide sufficient information that the reader can assess the methods used to generate the random allocation sequence and the likelihood of bias in group assignment". The Cochrane Handbook states

> "the starting point for an unbiased intervention study is the use of a mechanism that ensures that the same sorts of participants receive each intervention...If future assignments can be anticipated, either by predicting them or by knowing them, then selection bias can arise due to the selective enrolment and non-enrolment of participants into a study in the light of the upcoming intervention assignment."

In any proposed randomization sequence, there is risk that it can be either tampered with by someone involved in allocation (e.g. covertly breaking the sequence in order to assign the intervention to those seen as more needy) or simply inadvertently deterministic due to poor design (e.g. either by assigning treatment using a sequence that could be predicted by participants who would then selectively enroll or by deterministically assigning participants to groups by a rule relying on a nonrandom characteristic such as birth date), both of which can result in nonrandom treatment allocation and therefore biased treatment effect estimates. (Wood et al. 2008; K. Schulz et al. 2010)

Performance bias

Also known as the set of "Hawthorne" and "O. Henry" effects, performance bias is the tendency for participants to change their behavior or responses to questions because they are aware of being in a trial and of their treatment allocation. In many medical trials blinding of participants is used to minimize

this type of bias, as a participant unaware of allocation status is mechanically unable to act upon that knowledge. Discovery of allocation status in trials which were intended to be blinded has been linked to skewed results, a famous example of which is a 1975 study of the effects of Ascorbic Acid on the common cold, whose main result was that the (small) measured treatment effect was likely due to such bias. (Karlowski et al. 1975) The Cochrane Handbook states:

> "Lack of blinding of participants or healthcare providers could bias the results by affecting the *actual* outcomes of the participants in the trial. This may be due to a lack of expectations in a control group, or due to differential behaviours [sic] across intervention groups (for example, differential drop-out, differential cross-over to an alternative intervention, or differential administration of co-interventions) ... lack of blinding might also lead to bias caused by additional investigations or co-interventions regardless of the type of outcomes, if these occur differentially across intervention groups."

In the cases where blinding is impossible, it is essential to recognize and address concerns of bias resulting from knowledge of treatment allocation, as knowledge of differential treatment status is likely to be linked to differential subsequent outcome-related actions (e.g. seeking alternative treatments).

The likely direction of performance bias is ambiguous. On the one hand, the placebo effect is well known. A recent and salient example of this is a meta-analysis of studies of acupuncture treatment on back pain which showed that while acupuncture was superior to control interventions, (unblinded studies) it could not be proven to be superior to sham-interventions (blinded studies). (Ernst and White 1998) Conversely, in an RCT evaluating a medical intervention, if participants in the control group were aware of the intervention group treatment strategy, we might expect them to be more likely to seek

outside care than before as a result of said awareness. This differential care-seeking would introduce a systematic (downward) bias on the results. Risk of such bias is difficult to control for in many trials, particularly those in economics, as blinding is often impossible and the counterfactual - "what would the group have done if they had not been aware of their treatment allocation?" - cannot be answered. Nonetheless, CONSORT maintains that the possibility that knowledge of treatment allocation could skew behavior of the two groups differentially should be explicitly addressed in reports of RCTs in order to accurately assess the quality of data the trial provides.

The other face of this concern is that participants may respond differentially to questions if they know their treatment assignment. For example, in a study evaluating the impact of frequent nurse visits on morbidity from common cold, these nurse visits could induce the intervention group to pay more close attention to potential morbidity than the controls, regardless of whether actual morbidity is affected. The assessor should therefore be concerned that the nature of the intervention might skew the data collected, particularly if it is self-reported or otherwise subjective.[2] Lack of blinding has been linked to 30% exaggerated treatment effect estimates in a meta-analysis of studies with subjective outcomes. (Wood et al. 2008)

Detection bias

As with performance bias, detection bias is concerned primarily with blinding. In this case, however, the concern is about those collecting the data, not those providing it. The Cochrane Handbook warns that if "outcome assessors are aware of assignments, bias could be introduced into *assessments* of outcome, depending on who measures the outcomes." (Higgins, Green, and Collaboration 2008) (italics original) A trial evaluating the impact of blinding data assessors on measured treatment effect showed that

[2] Though all outcome assessments can be influenced by lack of blinding, there are particular risks of bias with more subjective outcomes (e.g. pain or number of days with a common cold). It is therefore recommended in these instruments to consider how subjective an outcome is when considering blinding.

preconceptions of treatment efficacy and placebo effects can have similar effects on data collectors and assessors as they do on participants. (Noseworthy et al. 1994) The CONSORT Statement adds that "unblinded data collectors may differentially assess outcomes (such as frequency or timing), repeat measurements of abnormal findings, or provide encouragement during performance testing. Unblinded outcome adjudicators may differentially assess subjective outcomes." (Moher et al. 2010) Evidence of detection bias has also been found in a trial in which ill patients performed a walking test with and without encouragement from the data collector. Encouragement alone was shown to improve time and distance walked by around 15%. (Guyatt et al. 1984) Unblinded trials which are not scrupulous in identifying and training data collectors to avoid these problems (and reporting efforts to this effect) are therefore at higher risk of detection bias.

Attrition bias

Attrition bias refers to a systematic loss of participants over the course of a trial in a manner that makes the final sample differ substantially from the initial sample, potentially destroying the comparability of treatment groups obtained by randomization. At the heart of these concerns is a practice known as "intention-to-treat analysis", wherein all participants who have gone through randomization should be included in the final analysis. One way of perceiving this concern is as an extension of the concerns outlined in the selection bias section taken forward to the execution and completion of the trial. Loss of participants can come from any number of reasons: drop-out, missing data, refusal to respond, death, or any exclusion rules applied after randomization. As explained in an article discussing this bias: "any analysis which omits patients is open to bias because it no longer compares the groups as randomised [sic]." (Lewis and Machin 1993)

One particularly salient example of post-hoc exclusion creating bias is the Anturane Trials, wherein the authors excluded those participants who died during the course of the trial, despite the fact

that mortality rates differed highly between control and intervention groups. The initial article from these trials showed a significant effect of the drug, but subsequent analyses which included participants according to randomization status failed to reject the null of no treatment effect. (Temple and Pledger 1980)

Reporting bias

Perhaps the most insidious of the problems facing those reading the reports of RCTs, reporting bias is the concern that authors present only a subset of their analyses and, as a result, the reader is left with only an incomplete and often a skewed understanding of the results. The more serious risk is that this bias will lead to many false positive conclusions about the efficacy of treatment and this, in turn, will lead to misinformed care or policy. The likelihood of this risk has been identified in a review of oncology articles published in two major medical journals, (Tannock 1996) and a more recent article confirmed this finding in three separate meta-analyses, finding that "statistically significant outcomes had a [sic] higher odds of being fully reported compared to non-significant outcomes (range of odds ratios: 2.2 to 4.7)." d (Dwan et al. 2008) A recent meta-analysis of studies on anthelminth therapy and treatment for incontinence found additional evidence that "more outcomes had been measured than were reported", and calculated that with a change in the assumptions about which outcomes the largest study chose to report, "the conclusions could easily be reversed." (Hutton et al. 2000)

To combat this problem, medical journals take two major steps. One, they require that a protocol be registered with a central, third-party database before the study begins. The protocol documents the plan for conduct of the trial, the intended sample size, and the analyses that the trialists will undertake at the end. This ensures continuity in the conduct of the trial, as any post-hoc changes that are made, potentially in favor of presenting more interesting results, would contradict the publicly

available plan for action. Upon consideration for publication, journal editors and peer reviewers can use the protocol to check for this.

The second tool is the definition of a "primary" analysis. The construction of t-test or similar comparison of means with a 95% confidence interval is such that conducting 20 such analyses will on average yield at least one "significant" result by virtue of chance alone. To prevent authors from running analyses ad infinitum and publishing only those which are significant, both the protocol and subsequent report of the article must report which analysis is given highest credence, also called the "primary outcome."

In this process, additional labels of "secondary" and "exploratory" outcomes are required to be assigned to the remaining presented results. This allows the reader to differentiate between analyses that the authors planned before the study and analyses which were conducted after the data was collected, thus with the benefit of being able to "peek" at the data and discern which relationships are significant. Exploratory analyses should be treated as hypothesis-forming, rather than hypothesis-answering, results, as there is a high risk of false-positive results in such analyses. (Oxman and Guyatt 1992; Yusuf et al. 1991; Assmann et al. 2000)

Sample size bias

Sample size calculations are a deceptively simple tool to minimize bias – an insufficiently large sample size can lead to imprecise estimation, biasing the results downward. CONSORT describes the risk of attenuation from sample size bias:

> "Reports of studies with small samples frequently include the erroneous conclusion that the intervention groups do not differ, when in fact too few patients were studied to make such a claim. Reviews of published trials have consistently found that a high

> proportion of trials have low power to detect clinically meaningful treatment effects. In reality, small but clinically meaningful true differences are much more likely than large differences to exist, but large trials are required to detect them."

A recent study of the issue also finds that trials with inadequate power have a high false-negative error rate and are implicated as a source of publication bias. (Dwan et al. 2008) Disentangling this result, two other studies found that small sample sizes were likely to overstate the effect size because of the heightened influence of outliers in these cases. (Moore, Gavaghan, et al. 1998; Moore, Tramer, et al. 1998) To guard against these problems, both CONSORT and Cochrane expect trialists to conduct sample size calculations before collecting any data and report them in trial publications.

As mentioned earlier, adherence to the CONSORT guidelines is now required by many journal editors for publication (K. Schulz et al. 2010) and articles which are successfully published in peer reviewed journals are again scrutinized by Cochrane Collaboration contributors during the conduct of systematic reviews. The result of this repeated scrutiny is a reduction in the problems described above. (Plint et al. 2006)

Economics trials and our motivation

As discussed in the introduction, academics in pure (as opposed to medical) economics departments have witnessed a surge in the use and popularity of RCTs in the last ten years. In the past half-century, governments and aid groups often relied on narrative accounts and "very rarely any firm evidence" to decide which programs to implement. (Pritchett 2002) During this period, evaluations of these projects are also claimed to have been performed on an ad-hoc basis and with very little statistical care. (Banerjee 2007) Perhaps as a result, economics RCTs were rapidly incorporated into the policy and

academic dialogues concerned with addressing this lacuna and are enjoying a stratospheric ascent similar to that of medical RCTs in the last half of the 20th century. (Banerjee 2007; Parker 2010)

Review of the bibliographies of peer-reviewed journal articles reporting economics RCTs reveals that many of the trials conducted to date have not explicitly drawn from the health literature on how to minimize bias in such experiments in the ways discussed above. As a result, we are concerned that economics trials unnecessarily risk stumbling into the same pitfalls which have plagued medical trials for the past sixty years.

In the section that follows, we describe the development and application of the grid, an instrument which uses the insights from the literature cited above as its main source. We are eager to acknowledge that the goals of economics trials are not identical to those of phase III medical trials and that it is an important question to ask how the metrics used to evaluate them should also differ. In light of this concern, the grid does not perfectly mirror the CONSORT Statement or Cochrane Handbook. Rather, it incorporates those suggestions which seem most appropriate to economics and excludes others which are either inappropriate for most economics trials (allocation concealment) or insufficiently objective (generalizability).

As for the criteria which remain, we contend that there are two justifications for applying them to the economics literature. One is that we see this as a $100 bill lying on the ground. The medical literature has carefully identified a set of well-defined concerns and shown that lack of attention to them yields bias in treatment effect estimates. There seems little reason not to use this advice. Perhaps more controversially, we recognize that many RCTs are used to inform development policy and domestic social policy in the developed world. To the extent that these policies affect a large number of lives in the developing and developed worlds, we think the same standards should be applied to these policy decisions as are applied to the decision whether to approve a wide array of non-vital pharmaceuticals in the US such as prescription anti-balding medicine.

Section III: Methodology

In this paper, we hope to answer the following research question: are the recent reports of RCTs in economics providing readers with sufficient information to assess the quality of evidence provided by the experiment (henceforth: are they adequately reporting how the trials were conducted) and is there evidence that authors take the necessary steps to minimize the risk of the biases that medical trialists have encountered? To answer this question, we developed a reporting and bias evaluation tool using a subset of the standards and guidelines set forth in CONSORT and the Cochrane Handbook. We then collected all economics articles published in a set of 50 major peer reviewed journals and, to evaluate the validity of our grid and to provide a benchmark for our ratings of articles in economics, we randomly selected an equal number of articles from peer reviewed journals in medicine. Finally, we applied our grid to both sets of articles. Below we describe our grid, our article selection process, and the assessment process itself.

The grid

To systematise the assessment of articles, we developed a grid which addresses each of the issues discussed in section II, provides leading questions to assist the assessor in assessment, and facilitates data collection. The full grid is given in Appendix 2. It is designed to facilitate and collect assessments of adequacy of reporting and risk of bias in terms of the six biases. There are 13 broad "issues" spread across the six biases, and many of these contain several smaller questions. The task of the assessor is to answer each question by putting either a "√" for yes or an "X" for no to the left of the question and, if at all possible, provide a page number or explanation in the comment and quote boxes to the right of the question to justify the assessment. The assessor then aggregates the assessments from questions to issues, and then aggregates from issues to an overall assessment for each of the six

biases, separately for adequacy of reporting and risk of bias using a simple rule: if the article fails on any issue in terms of adequacy of reporting, then it fails for the overall adequacy of reporting of that bias (and similarly for the assessment of low risk of bias). The motivation for this structure is that each type of bias is complex, comprising several different concerns, each of which must be addressed to minimize the risk of a given bias.

The result of this grading process was an assessment for each of the 13 issues and each of the 6 biases, whether the issue/bias was reported adequately, and whether or not there was low risk of bias associated with that issue/bias.

The studies

For this analysis, we collected a set of articles published in peer-reviewed journals in economics reporting the results of economics trials. The selection process was as follows:

1) Using the EconLit database, we searched for journal articles published between 2000 and 2009 that contained either the word randomized or randomization (or their alternative British spellings) in the title or abstract. A search conducted on July 6th, 2010 generated 527 results.
2) From these results, we further limited eligibility with two criteria:
    a. The first eligibility criterion was that an article had to report the results of a prospectively randomized study. This condition was incorporated in light of the fact that we are evaluating study design and so it would be inappropriate to include studies not specifically designed as trials (e.g. public lotteries or other natural experiments).
    b. To limit heterogeneity of study quality, we further restricted eligibility to articles published in the top 50 journals as rated by journal impact within economics, taken

from a Boston Fed working paper which ranks economics journals. (Kodrzycki and Yu 2005)

In total, this yielded 28 articles published between 2001 and 2009. A full list is provided in Appendix 1.

We randomly selected an equal number of articles reporting RCTs published in three of the top peer-reviewed medical journals for grading. This served two purposes – one, we wanted to use these grades to see whether our instrument and grading process were on target. If the instrument failed these articles on most accounts, we would be worried that it might be too strict as adherence to the CONSORT standards was required by most medical journal editors during this period. Two, we wanted to provide a benchmark for how the "gold standard" in medicine would fare according to our standards. We drew our sample such that in each year with at least one eligible article in economics, there were selected an equal number of articles in medicine as there were eligible articles in economics. We chose to draw this sample of articles in medicine from the top three medical journals as classified by the Thompson *Journal and Citation Reports*' impact factor in general and internal medicine as of July 6th, 2010. (Thompson Reuters 2010) These journals are *The Lancet, The Journal of the American Medical Association,* and *The New England Journal of Medicine*. The decision to only consider articles from these three journals was made with two motives: one, for ease of processing, as there are literally thousands of RCT reports published each year and restricting the journals to these three still left us with approximately 350 each year and, two, in order to see how our grid fared evaluating the "gold standard" in medicine.

To obtain the medical RCT article sample, we followed the following process

1) We searched Pubmed (a database similar to Econlit indexing medical journals and their articles) for all articles reporting clinical trials in these three journals in years when there was also an eligible economics article published (all years in our range save 2000 and 2002).

2) From this list, we then randomly selected a number of articles in a given year equal to the number of eligible articles in economics in that year. Randomization was performed by ordering the journal articles as they appeared in the search, assigning each article a random number between 0 and 1 using a random number generator, and then sorting the articles in ascending order by the magnitude of the randomly assigned number, selecting the first x articles required to achieve balance between the two fields.
3) We excluded Phase I and II trials in medicine as their methods, goals and sample size considerations are significantly different from Phase III trials, which, similar to the economics trials we are concerned with, are more often used to inform policy. The final list of these papers is also given in Appendix 1.

In both medicine and economics, if a trial generated more than one eligible publication, the article published earliest was selected. Other associated articles were used to provide additional information for evaluation of the main article only.

The assessment process

Both authors first read each article and assessed the adequacy of reporting and risk of bias using the grid individually. For each article, we then discussed our assessments. Any disagreements were resolved through deliberation, the result of which is the final assessment of each study, presented in section IV. This method of individual grading followed by deliberation was adopted in light of the difficulty met in medicine during efforts to conduct similar studies which elected to use independent grading without deliberation. While the latter type of grading potentially provides better evidence in favor of the internal validity of the grid, the rate of agreement between graders in such processes is

often low. (Clark et al. 1999) We chose a deliberation process instead to ensure that the results from grading, which is a long and tedious process, were as reliable as possible.

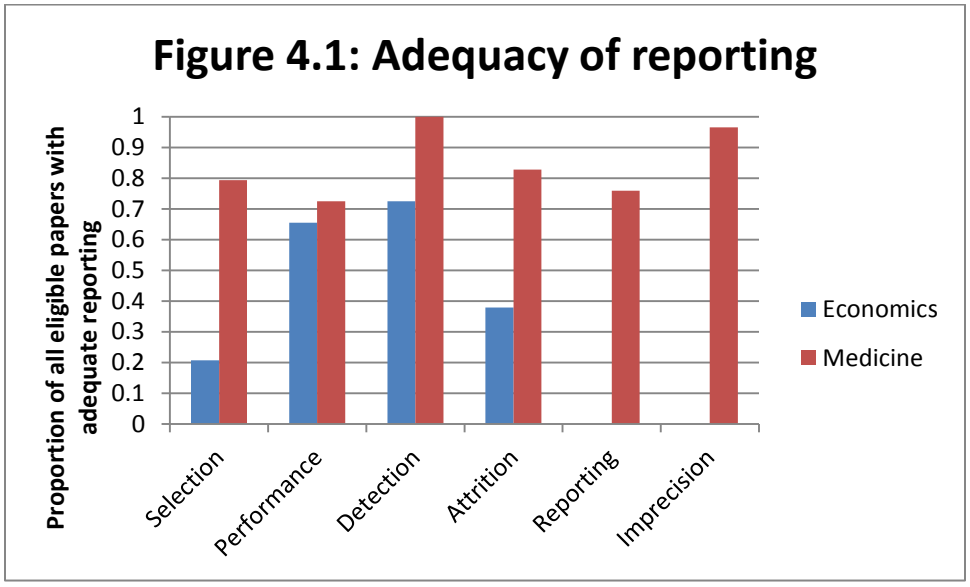
In the analysis on risk of bias that follows, we group inadequacy of reporting (and therefore unclear risk of bias) with high risk of bias. While this is not ideal, unclear risk of bias sheds similar, if not as severe, doubts on the conclusions of the study in question. We draw this method from the landmark meta-analysis assessing study quality in medicine. (K. F. Schulz et al. 1995) Furthermore, we do not aggregate the individual scores to create an overall study-level score, as each section represents a separate concern, again following the lead of meta-analyses in medicine. (Spiegelhalter and Best 2003) As the issues in our analysis are diverse, both the effect size and direction of bias are likely to be different for each individual bias, if not issue.

Section IV: Analysis

In this section we compare our assessments of published articles in economics and medicine, in terms of adequacy of reporting and risk of bias. We find that the economics literature reports on the majority of these risks irregularly - for three of the six issues, less than 50 percent of the articles collected report adequately, and for no type of bias do more than three fourths of the articles report adequately. The pattern is largely similar, with uniformly lower rates of passing, for our assessments of risk of bias in economics articles. Though much of this relationship is mechanical, even among the subset of articles in which reporting is adequate there are many cases in which there is significant risk of bias. For two of the six biases, all articles in economics that we include fail to report adequately and cannot be assessed as having low risk of bias. The medical literature, as expected, does much better, though for no bias do 100 percent of the articles report adequately or have low risk of bias.

Below, we show summary statistics of our assessments and then provide selected examples of concerns from the economics articles. Figure 4.1 shows the overall performance of eligible articles in the

two disciplines in terms of adequacy of reporting. Figure 4.2 shows the same comparison for our analysis of risk of bias. Figure 4.3 displays the proportion of articles in economics which were assessed as not having low risk of bias whose assessment was due to inadequate reporting. Table 4.1 gives the overall percentage of papers with adequate reporting and low risk of bias for each of the six biases. It also shows the raw difference between performance of articles in the two disciplines and the p-value results of a simple Student's T test for mean equality for reporting and risk of each bias. In the rest of this section, we describe the bias-specific performance of economics articles and explain some of the major driving factors behind this performance.



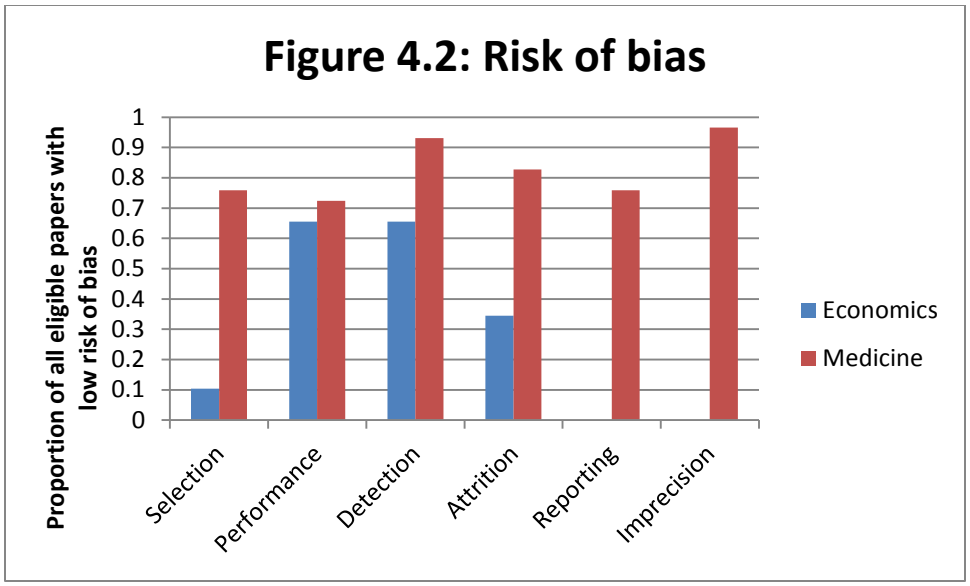


Figure 4.2: Risk of bias
Proportion of all eligible papers with low risk of bias
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Selection
Performance
Detection
Attrition
Reporting
Imprecision
Economics
Medicine

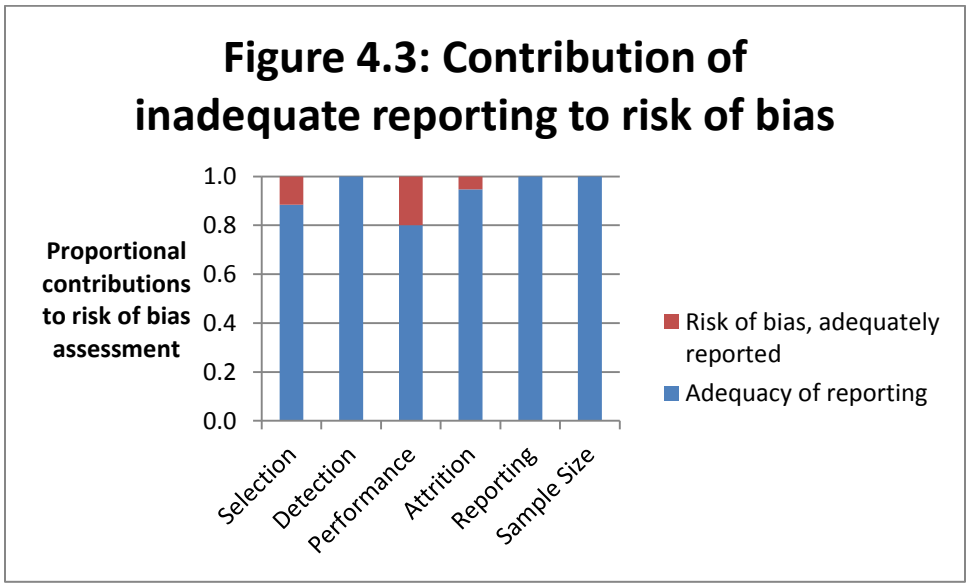
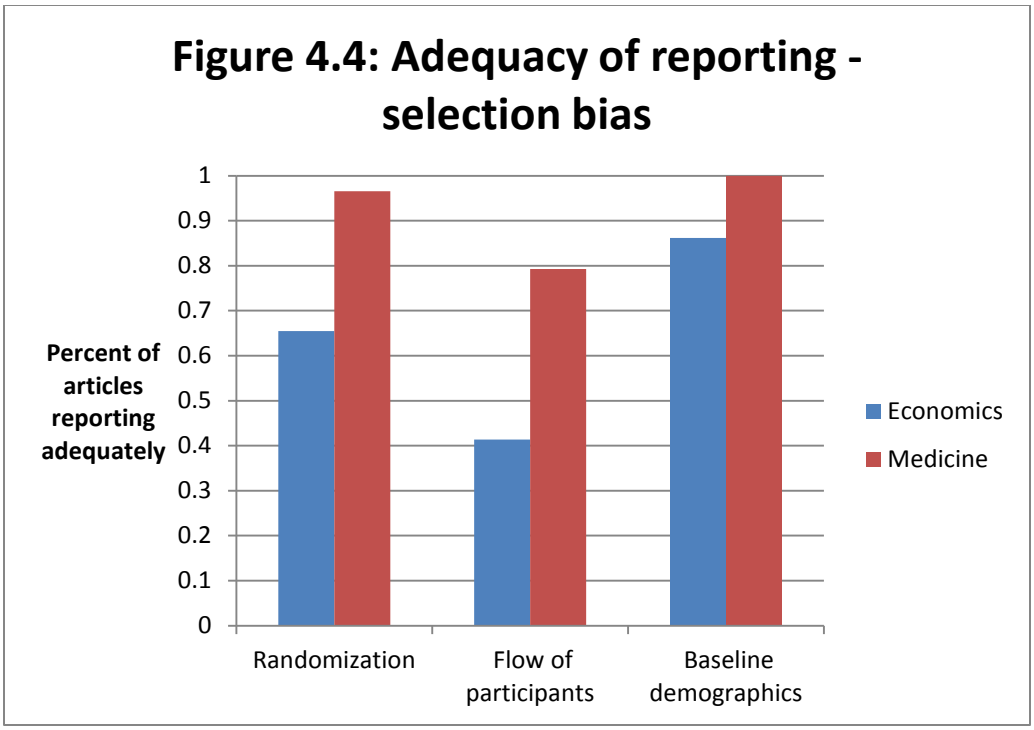


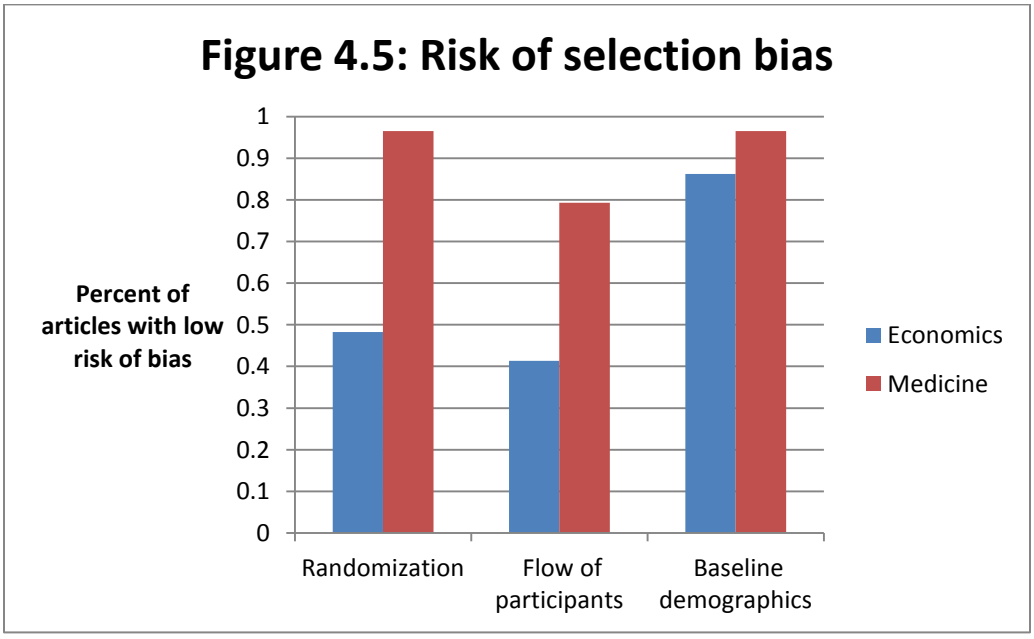
Figure 4.3: Contribution of inadequate reporting to risk of bias
Proportional contributions to risk of bias assessment
1.0
0.8
0.6
0.4
0.2
0.0
Selection
Detection
Performance
Attrition
Reporting
Sample Size
Risk of bias, adequately reported
Adequacy of reporting

Table 4.1 – Performance of articles by issue and discipline

Bias	Issue	Economics articles passing	Medical articles passing	T-test p-value
Selection	Reporting	20.7%	79.3%	0.000
Selection	Risk of bias	10.3%	75.9%	0.000
Performance	Reporting	65.5%	72.4%	0.578
Performance	Risk of bias	65.5%	72.4%	0.578
Detection	Reporting	72.4%	100.0%	0.003
Detection	Risk of bias	65.5%	93.1%	0.010
Attrition	Reporting	37.9%	82.8%	0.000
Attrition	Risk of bias	34.5%	82.8%	0.000
Reporting	Reporting	0.0%	75.9%	0.000
Reporting	Risk of bias	0.0%	75.9%	0.000
Imprecision	Reporting	0.0%	96.6%	0.000
Imprecision	Risk of bias	0.0%	96.6%	0.000

Selection bias: Only six of the 28 eligible economics articles (21%) passed the reporting criteria for selection bias. For reference, almost 80% of eligible medical articles did so. Figures 4.4 and 4.5 present performance of both economics and medical articles on the three issues comprising this bias – randomization, flow of participants, and baseline demographics.





Ten of the 28 articles in economics did not pass because they provided insufficient details on the process used to randomize, an ambiguity which leaves doubt as to whether the randomization

processes used could have been deterministic or that the administrator could have corrupted the sequence. Five articles did report their means of randomization, but used clearly deterministic methods (for example, an alphabetic algorithm in one case and sorting by date of employment commencement in another) to assign treatment. Lack of information about the flow of potential participants in the trial was another major flaw in articles in economics. In 17 of the eligible articles published in economics journals, the numbers of participants screened for eligibility and excluded before and after eligibility were not given.

Performance bias: Again, 10 economics papers reported inadequately in terms of performance bias and an equal amount had high risk of bias. In most medical trials, this problem is often avoided by administering placebos to controls and thus blinding participants to which treatment group they have been assigned to. In some instances, this is impossible, but when blinding is not feasible, the medical literature (and our grid) requires that the authors of the study discuss the potential for such bias and demonstrate that it was not in fact a risk. The economics papers which failed on these criteria almost uniformly neglected to address this concern. It is important to note that we did not fail papers for not blinding – rather, a paper did not pass on adequacy of reporting if there was apparent risk of performance bias (e.g. alternative care seeking as a result of knowledge of treatment status) which was not discussed. In an article which evaluated a program which gave cash transfers conditional on school enrolment, for example, there is a clear concern that participants assigned to the control group would change their behavior (by waiting to send children to school, for example, until the program was rolled out to all households) in light of their knowledge of their and others' treatment status. There was no mention of this concern in the article in question.

Detection bias: shortcomings in terms of detection bias had to do with the identity of data collectors and the nature of data. Eight of the 28 economics articles failed on reporting and 10 of 28 on risk of this bias. Many of these trials collected data with individuals who may have had incentive to skew the data in favor of the intervention. One article explicitly mentioned using data collectors who were employed by the same company which administered the intervention. Several others neglected to say who collected the data, leaving doubt as to whether a similar conflict of interest could have biased the results.

Attrition bias: there are two interlinked concerns here – one is that participants dropped out during the course of the trial in a way that would destroy the balance between treatment groups achieved by randomization. The other concern is that the analyses run do not follow the "intent to treat" principle, which stipulates that all randomized participants be included in the final analysis. Only ten of the economics articles passed this criterion. Seventeen did not discuss exclusion of participants in the final analysis and almost all of these had widely varying numbers of observations in different versions of the same analysis, suggesting that selective exclusion of observations did in fact take place. Less than half of the articles we collected mentioned the intent to treat principle by name and, among those that did, several neglected to follow it. Many of these articles excluded groups of participants because they did not follow the protocol, and one paper threw out the second of two years of data collected because of contamination. While these concerns do not definitively show bias, they leave open the possibility for bias from attrition, an ambiguity that has been associated with exaggerated results in medical trials.

Reporting bias: No economics paper was adequately reported in terms of reporting bias, and therefore none could be assessed as having low risk of bias in this category. Our assessments attest to two phenomena. The first and foremost is the lack of pre-specification of endpoints and ad-ante publication of a study protocol. As described in Section II, pre-specifying a primary endpoint in a protocol published

before the trial begins ties the authors' hands and forces them to present only one analysis as the "primary" finding. All other analyses are meant to be specified as either secondary or ad-hoc, thus addressing the concern that a selectively chosen subset of all conducted analyses are presented and given more than the appropriate weight in the discussion of results. No economics paper attested to do this, almost certainly because this is not required practice for publication in peer-reviewed economics journals. We are aware of the fact that writing a protocol and registering it is now required by groups such as JPAL, however this was not mentioned in any of the studies and no links or references to protocols were provided.

The other issue at hand in reporting bias is that of even-handedness in presentation of results. Nearly half of the economics papers were not deemed to interpret their results adequately. Broken down, this is largely due to the fact that the articles did not mention whether there were any limitations in their methods nor did they condition their interpretation of the strength of their results in light of the many comparisons that they presented. This is perhaps the most controversial element of the medical literature which we propose to apply to economics. Neither the medical literature on the matter nor the authors of this paper wish to say that results generated after the trial concludes and shown to be robust under several specifications should not be used in the formation of policy. Rather, the medical literature suggests that the burden of proof for such analyses should be much higher than for pre-specified hypotheses.

Imprecision: here, very simply, there was no indication that any eligible economics article performed an ad ante sample size calculation. We are almost certain that some did, (Parker 2010) but none were reported. Though contacting authors to solicit such information was a possibility, there is evidence that doing so would lead to biased responses (Haahr and Hróbjartsson 2006) and our rule tying inadequacy of reporting to risk of bias was applied.

Section V: Ways forward

We have presented evidence that the standards for reporting and conduct of RCTs in economics are not in line with medical standards on conduct and reporting and as a result, trials in economics are at risk of bias in their analyses. Our work draws on a body of medical literature which has linked poor trial design, conduct, and reporting to exaggerated estimates of treatment effects. (Moher et al. 1998; K. Schulz et al. 2010; K. Schulz et al. 2010) The identification of these shortcomings led to the systems of standards now used by medical trialists and journal editors which we draw upon for our grid. The establishment and acceptance of these standards in medicine has, in turn, led to an increase in the quality of articles reporting the results of trials. (Plint et al. 2006)

Similar issues in the economics literature have been brought to light in the past few years. A recent exchange between Deaton and Imbens touches on many of these concerns. (Imbens 2009; Deaton 2009) Despite their divergent views, the two authors agree on the fact that poor conduct of RCTs can bias interpretation. A more thorough description of these concerns and other more practical problems of RCT implementation and interpretation is given in Duflo, Glennerster, and Kremer's article on how to conduct RCTs. (Duflo, Glennerster, and Kremer 2007) We are aware that registration of a protocol and analysis plan is now common practice at JPAL and other centers. Still, there seems to be no consensus on how to run an RCT in the social sciences and our analysis suggests that economists have not adopted many of the tools that medical trialists use for minimizing the risk of certain biases in their reports.

To ensure that the quality of evidence provided in economics articles reporting the results of RCTs is as high as possible, we propose that a system of reporting standards be established in economics similar to the CONSORT guidelines widely accepted in the medical literature. These standards would guide authors a tool to use on three fronts: one, in writing scholarly articles reporting the results of RCTs

for publication in peer-reviewed journals, two, in the initial design of the studies themselves, and three, in performing meta-analyses and critical reviews such as this article. The crux of the argument in favor of such standards is twofold: one, that providing this information in trial reports enables readers to assess the quality of the evidence provided in each article, and two, that enforcing such standards encourages careful conduct of trials as well as thorough reporting.

In terms of implementation, the standards for trials in economics would necessarily differ substantially from those of medicine, (perhaps in the nature of the requirement of pre-specification of endpoints, for example) and the contents of such a system would have to come from a consensus among economists on what constitutes good practice as well as which data are necessary to assess trial quality. Duflo, Glennerster and Kremer's article outlines several issues that should be included in any set of guidelines, (Duflo, Glennerster, and Kremer 2007) but their treatment of the issues is not exhaustive. For our part, we believe that at the very least, the following issues from CONSORT should be part of any set of guidelines for RCT design and reporting: a CONSORT-style diagram of flow of participants, a trial protocol registration system, which would include prespecifying a primary analysis and providing explicit, ad-ante sample size calculations, and insistence on the intent-to-treat principle for the primary analysis.

We also recognize that this is a field ripe for more analysis. Productive avenues of inquiry include mathematical simulation of the different types of biases to estimate how much the treatment effects in the literature to date should be discounted, investigation of publication bias in RCTs, and constructing a taxonomy of phases for trials in economics to help us know better when and how to apply the lessons from bias in medical trials. Additionally, though our initial investigation engaged with questions of external validity as well as internal, we have restricted our discussion here to internal validity to make our message more concise. This is unfortunate, as external validity is arguably of similar importance and there is a similarly rich literature on how to assess this in reports of RCTs. (Rothwell

2006) Each of these, however, is beyond the scope of this paper and we leave their pursuit to future research.

Lastly, we would like to mention a few of the weaknesses of our study. Our grading task was a long and tedious one and almost certainly not without some human error. An increase in the number of evaluators for each paper would improve the reliability of our results. That said, the differences we find between the two sets of RCT reports are so stark that we think it unlikely to be solely explicable as measurement error. Also, further analyses, such as examination of trends in the quality of economics trials over time, or between journals, promise to be fruitful. As time goes on, more reports will be published and there is likely to be a more adequate sample size to make such comparisons. The number of eligible studies published during the period we cover in this study was small enough to make such an exercise appear less than worthwhile.

Section VI: Conclusion

In this article, we have identified and discussed the potential for bias in the reports of randomized controlled trials in economics. From two of the main bias identification and minimization tools used by the medical literature, we crafted an evaluation tool, which we call the grid, to evaluate the adequacy of reporting and risk of six major biases in RCTs in economics. We evaluated a set of articles reporting the results of RCTs from 50 top economics journals and found that these articles performed poorly both in terms of providing the reader adequate information with which to assess the quality of the evidence provided by the study, and in terms of minimizing the risk of these six types of bias which have been associated with exaggerated treatment effects. We concluded by suggesting that the field of economics develop and adopt a set of reporting guidelines both to require the same degree of clarity and precision in the reports of RCTs that is demanded in medicine and to serve as a quality assessment tool to evaluate results that are published.

There are two main contributions of our analysis: methodological and empirical. In terms of methodology, we have discussed the nature of a set of biases and problems we believe RCTs are particularly prone to, catalogued the evidence of such problems skewing results in the medical literature, and provided a tool which can be used both to evaluate risk of bias in reports of RCTs as well as to assist in the design of future RCTs. Empirically, we showed that the reports of trials in economics published between 2000 and 2009 inadequately reported the risks of these bias according to the standards we derived from the medical literature, and that the design and implementation of many of these trials suggests they have made mistakes similar to those made in the past in the medical literature. Both findings suggest problems which have been associated with exaggerated treatment effects in the medical literature and raise serious concerns about the strength of the conclusions reached in some of the eligible studies in economics.

Going forward, we hope that our study will lead to the establishment and acceptance of a set of standards for reporting RCTs that will minimize these biases in published reports of RCTs in the economics literature and will help readers to assess the quality of evidence provided in these reports. We hope it will also lead to increased efforts by trialists themselves to avoid these pitfalls in the design, execution, and analysis of their trials. Such efforts would lead to higher quality policy advice and, we hope, the implementation of policy closer to the optimal.

References:

1. Assmann, S. F, S. J Pocock, L. E Enos, and L. E Kasten. 2000. “Subgroup Analysis and Other (mis) Uses of Baseline Data in Clinical Trials.” *The Lancet* 355 (9209): 1064–1069.
2. Banerjee, A. V. 2007. *Making Aid Work*. The MIT Press.
3. Bell, J. A. 1948. “PERTUSSIS IMMUNIZATION Use of Two Doses of an Alum-Precipitated Mixture of Diphtheria Toxoid and Pertussis Vaccine.” *Journal of the American Medical Association* 137 (15): 1276–1281.
4. Clark, H. D, G. A Wells, C. Huët, F. A McAlister, L. R Salmi, D. Fergusson, and A. Laupacis. 1999. “Assessing the Quality of Randomized Trials: Reliability of the Jadad Scale.” *Controlled Clinical Trials* 20 (5): 448–452.
5. Collier, Roger. 2009. “Legumes, Lemons and Streptomycin: A Short History of the Clinical Trial.” *CMAJ : Canadian Medical Association Journal* 180 (1) (January 6): 23–24. doi:10.1503/cmaj.081879.
6. Deaton, Angus S. 2009. “Instruments of Development: Randomization in the Tropics, and the Search for the Elusive Keys to Economic Development.” *National Bureau of Economic Research Working Paper Series* No. 14690. http://www.nber.org/papers/w14690.
7. Duflo, E., R. Glennerster, and M. Kremer. 2007. “Using Randomization in Development Economics Research: A Toolkit.” *Handbook of Development Economics* 4: 3895–3962.
8. Dwan, K., D. G Altman, J. A Arnaiz, J. Bloom, A. W Chan, E. Cronin, E. Decullier, et al. 2008. “Systematic Review of the Empirical Evidence of Study Publication Bias and Outcome Reporting Bias.” *PLoS One* 3 (8): e3081.
9. Ernst, E., and A. R White. 1998. “Acupuncture for Back Pain: a Meta-analysis of Randomized Controlled Trials.” *Archives of Internal Medicine* 158 (20): 2235.

10. Gluud, L. L. 2006. “Bias in Clinical Intervention Research.” *American Journal of Epidemiology* 163 (6): 493–501.
11. Guyatt, G. H., S. O. Pugsley, M. J. Sullivan, P. J. Thompson, L. Berman, N. L. Jones, E. L. Fallen, and D. W. Taylor. 1984. “Effect of Encouragement on Walking Test Performance.” *Thorax* 39 (11): 818–822.
12. Haahr, M. T, and A. Hróbjartsson. 2006. “Who Is Blinded in Randomized Clinical Trials?” *The Cochrane Collaboration Methods Groups Newsletter* 3: 14.
13. Higgins, J. P.T, S. Green, and C. Collaboration. 2008. *Cochrane Handbook for Systematic Reviews of Interventions*. Vol. 5. September. Wiley Online Library.
14. Hutton, J. L, Paula R Williamson, J. L Hutton, and Paula R Williamson. 2000. “Bias in Meta-analysis Due to Outcome Variable Selection Within Studies, Bias in Meta-analysis Due to Outcome Variable Selection Within Studies.” *Journal of the Royal Statistical Society: Series C (Applied Statistics), Journal of the Royal Statistical Society: Series C (Applied Statistics)* 49, 49 (3, 3) (January 1): 359, 359–370, 370. doi:10.1111/1467-9876.00197, 10.1111/1467-9876.00197.
15. Imbens, G. W. 2009. *Better LATE Than Nothing: Some Comments on Deaton (2009) and Heckman and Urzua (2009)*. National Bureau of Economic Research.
16. Ioannidis, J. P.A, A. B Haidich, M. Pappa, N. Pantazis, S. I Kokori, M. G Tektonidou, D. G Contopoulos-Ioannidis, and J. Lau. 2001. “Comparison of Evidence of Treatment Effects in Randomized and Nonrandomized Studies.” *JAMA: The Journal of the American Medical Association* 286 (7): 821–830.
17. Jüni, P., A. Witschi, R. Bloch, and M. Egger. 1999. “The Hazards of Scoring the Quality of Clinical Trials for Meta-analysis.” *JAMA: The Journal of the American Medical Association* 282 (11): 1054–1060.

18. Karlowski, T. R, T. C Chalmers, L. D Frenkel, A. Z Kapikian, T. L Lewis, and J. M Lynch. 1975. "Ascorbic Acid for the Common Cold." *JAMA: The Journal of the American Medical Association* 231 (10): 1038–1042.
19. Kodrzycki, Y., and P. Yu. 2005. "New Approaches to Ranking Economics Journals."
20. Lewis, J. A., and D. Machin. 1993. "Intention to Treat–who Should Use ITT?" *British Journal of Cancer* 68 (4): 647.
21. Meldrum, M. L. 2000. "A Brief History of the Randomized Controlled Trial: From Oranges and Lemons to the Gold Standard." *Hematology/oncology Clinics of North America* 14 (4): 745–760.
22. Moher, D., S. Hopewell, K. F. Schulz, V. Montori, P. C. Gøtzsche, P. J. Devereaux, D. Elbourne, M. Egger, and D. G. Altman. 2010. "Consolidated Standards of Reporting Trials Group.. CONSORT 2010 Explanation and Elaboration: Updated Guidelines for Reporting Parallel Group Randomised Trials." *Journal of Clinical Epidemiology* 63 (8): e1–37.
23. Moher, D., B. Pham, A. Jones, D. J Cook, A. R Jadad, M. Moher, P. Tugwell, and T. P Klassen. 1998. "Does Quality of Reports of Randomised Trials Affect Estimates of Intervention Efficacy Reported in Meta-analyses?" *The Lancet* 352 (9128): 609–613.
24. Moore, R. A., D. Gavaghan, M. R. Tramer, S. L. Collins, and H. J. McQuay. 1998. "Size Is Everything–large Amounts of Information Are Needed to Overcome Random Effects in Estimating Direction and Magnitude of Treatment Effects." *Pain* 78 (3): 209–216.
25. Moore, R. A., M. R. Tramer, D. Carroll, P. J. Wiffen, and H. J. McQuay. 1998. "Quantitive Systematic Review of Topically Applied Non-steroidal Anti-inflammatory Drugs." *BMJ: British Medical Journal* 316 (7128): 333–338.
26. Noseworthy, J. H, G. C Ebers, M. K Vandervoort, R. E. Farquhar, E. Yetisir, and R. Roberts. 1994. "The Impact of Blinding on the Results of a Randomized, Placebo-controlled Multiple Sclerosis Clinical Trial." *Neurology* 44 (1): 16–16.

27. Oxman, A. D, and G. H Guyatt. 1992. “A Consumer’s Guide to Subgroup Analyses.” *Annals of Internal Medicine* 116 (1): 78–84.
28. Parker, I. 2010. “The Poverty Lab: Transforming Development Economics, One Experiment at a Time.” *New Yorker* 17: 79–89.
29. Plint, A. C, D. Moher, A. Morrison, K. Schulz, D. G Altman, C. Hill, I. Gaboury, and others. 2006. “Does the CONSORT Checklist Improve the Quality of Reports of Randomised Controlled Trials? A Systematic Review.” *Medical Journal of Australia* 185 (5): 263.
30. Pritchett, L. 2002. “It Pays to Be Ignorant: a Simple Political Economy of Rigorous Program Evaluation.” *The Journal of Policy Reform* 5 (4): 251–269.
31. Rothwell, P. M. 2006. “Factors That Can Affect the External Validity of Randomised Controlled Trials.” *PLoS Hub for Clinical Trials* 1 (1): e9.
32. Schulz, K., D. Altman, D. Moher, and others. 2010. “CONSORT 2010 Statement: Updated Guidelines for Reporting Parallel Group Randomised Trials.” *BMC Medicine* 8 (1): 18.
33. Schulz, K. F, I. Chalmers, R. J Hayes, and D. G Altman. 1995. “Empirical Evidence of Bias.” *JAMA: The Journal of the American Medical Association* 273 (5): 408–412.
34. Spiegelhalter, D. J, and N. G Best. 2003. “Bayesian Approaches to Multiple Sources of Evidence and Uncertainty in Complex Cost-effectiveness Modelling.” *Statistics in Medicine* 22 (23): 3687–3709.
35. Tannock, I. F. 1996. “False-positive Results in Clinical Trials: Multiple Significance Tests and the Problem of Unreported Comparisons.” *Journal of the National Cancer Institute* 88 (3-4): 206–207.
36. Temple, R., and G. W Pledger. 1980. “The FDA’s Critique of the Anturane Reinfarction Trial.” *New England Journal of Medicine* 303 (25): 1488–1492.
37. The Cochrane Collaboration. 2010. “The Cochrane Collaboration, Home - The Cochrane Library.” http://www.thecochranelibrary.com/view/0/index.html.

38. Thompson Reuters. 2010. “Thompson Reuters, ISI Web of Knowledge Journal Citation Reports for Medicine, General & Internal.” http://admin-apps.isiknowledge.com/JCR/JCR.
39. Twyman, Richard. 2004. “A Brief History of Clinical Trials.” http://genome.wellcome.ac.uk/doc_WTD020948.html.
40. Vader, J. P. 1998. “Randomised Controlled Trials: A User’s Guide.” *BMJ: British Medical Journal* 317 (7167): 1258.
41. Wood, L., M. Egger, L. L Gluud, K. F Schulz, P. Jüni, D. G Altman, C. Gluud, R. M Martin, A. J.G Wood, and J. A.C Sterne. 2008. “Empirical Evidence of Bias in Treatment Effect Estimates in Controlled Trials with Different Interventions and Outcomes: Meta-epidemiological Study.” *BMJ: British Medical Journal* 336 (7644): 601–605.
42. Yusuf, S., J. Wittes, J. Probstfield, and H. A Tyroler. 1991. “Analysis and Interpretation of Treatment Effects in Subgroups of Patients in Randomized Clinical Trials.” *JAMA: The Journal of the American Medical Association* 266 (1): 93–98.

Appendix 1: Articles evaluated in the analysis

Articles in economics				
No.	First Author	Journal	Year	Title
1	Thornton	American Economic Review	2008	The Demand for, and Impact of, Learning HIV Status
2	van den Berg	International Economic Review	2006	Counseling and Monitoring of Unemployed Workers: Theory and Evidence from a Controlled Social Experiment
3	Angrist	American Economic Review	2009	The Effects of High Stakes High School Achievement Awards: Evidence from a Randomized Trial
4	Ashenfelter	Journal of Econometrics	2005	Do Unemployment Insurance Recipients Actively Seek Work? Evidence from Randomized Trials in Four U.S. States
5	Ashraf	Quarterly Journal of Economics	2006	Tying Odysseus to the Mast: Evidence from a Commitment Savings Product in the Philippines
6	Banerjee	Quarterly Journal of Economics	2007	Remedying Education: Evidence from Two Randomized Experiments in India
7	Bjorkman	Quarterly Journal of Economics	2009	Power to the People: Evidence from a Randomized Field Experiment on Community-Based Monitoring in Uganda
8	Bobonis	Journal of Human Resources	2006	Anemia and School Participation
9	Cai	American Economic Review	2009	Observational Learning: Evidence from a Randomized Natural Field Experiment
10	de Mel	Quarterly Journal of Economics	2008	Returns to Capital in Microenterprises: Evidence from a Field Experiment
11	Duflo	Quarterly Journal of Economics	2006	Saving Incentives for Low- and Middle-Income Families: Evidence from a Field Experiment with H&R Block
12	Duflo	Quarterly Journal of Economics	2003	The Role of Information and Social Interactions in Retirement Plan Decisions: Evidence from a Randomized Experiment
13	Fehr	American Economic Review	2007	Do Workers Work More if Wages Are High? Evidence from a Randomized Field Experiment
14	Gine	Journal of Development Economics	2009	Insurance, Credit, and Technology Adoption: Field Experimental Evidence from Malawi
15	Glewwe	Journal of Development Economics	2004	Retrospective vs. Prospective Analyses of School Inputs: The Case of Flip Charts in Kenya
16	Harrison	Journal of Economic	2009	Risk Attitudes, Randomization to Treatment, and Self-Selection

		Behavior and Organization		into Experiments
17	Hu	Journal of Human Resources	2003	Marriage and Economic Incentives: Evidence from a Welfare Experiment
18	Karlan	American Economic Review	2008	Credit Elasticities in Less-Developed Economies: Implications for Microfinance,
19	Katz	Quarterly Journal of Economics	2001	Moving to Opportunity in Boston: Early Results of a Randomized Mobility Experiment
20	Kremer	Quarterly Journal of Economics	2007	The Illusion of Sustainability
21	Kremer	Review of Economics and Statistics	2009	Incentives to Learn
22	Michalopoulos	Journal of Public Economics	2005	When Financial Work Incentives Pay for Themselves: Evidence from a Randomized Social Experiment for Welfare Recipients
23	Miguel	Econometrica	2004	Worms: Identifying Impacts on Education and Health in the Presence of Treatment Externalities
24	Olken	Journal of Political Economy	2007	Monitoring Corruption: evidence from a Field Experiment in Indonesia
25	Pozo	American Economic Review	2006	Requiring a Math Skills Unit: Results of a Randomized Experiment
26	Rosholm	Journal of Applied Econometrics	2009	Is Labour Market Training a Curse for the Unemployed? Evidence from a Social Experiment
27	Schady	Economics Letters	2008	Are Cash Transfers Made to Women Spent Like Other Sources of Income?
28	Schultz	Journal of Development Economics	2004	School Subsidies for the Poor: Evaluating the Mexican Progresa Poverty Program

Articles in medicine				
1	Blankensteijn	New England Journal of Medicine	2005	Two-Year outcomes after Conventional or Endovascular Repair of Abdominal Aortic Aneurysms
2	Pichichero	Journal of the American Medical Association	2005	Combined Tetanus, Diphtheria, and 5-Component Pertussis Vaccine for use in Adolescents and Adults
3	Staessen	Journal of the American Medical Association	2004	Antihypertensive Treatment Based on Blood Pressure Measurement at Home or in the Physician's Office: A Randomized Controlled Trial

4	Tate	Journal of the American Medical Association	2003	Effects of Internet Behavioral Counseling on Weight Loss in Adults at Risk for Type 2 Diabetes: A Randomized Trial
5	Gorelick	Journal of the American Medical Association	2003	Aspirin and Ticlopidine for Prevention of Recurrent Stroke in Black Patients: A Randomized Trial
6	Riddler	New England Journal of Medicine	2008	Class-Sparing Regimens for Initial Treatment of HIV-1 Infection
7	Tardif	The Lancet	2008	Effects of succinobucol (AGI-1067) after an acute coronary syndrome: a randomised, double-blind, placebo controlled trial
8	Fergusson	New England Journal of Medicine	2008	A Comparison of Aprotinin and Lysine Analogues in High-Risk Cardiac Surgery
9	Scolnik	Journal of the American Medical Association	2006	Controlled Delivery of High vs Low Humidity vs Mist Therapy for Croup in Emergency Departments: A Randomized Controlled Trial
10	Tonetti	New England Journal of Medicine	2007	Treatment of Periodontitis and Endothelial Function
11	Papanikolaou	New England Journal of Medicine	2006	In Vitro Fertilization with Single Blastocyst-Stage versus Single Cleavage-Stage Embryos
12	Van den Berghe	New England Journal of Medicine	2006	Intensive Insulin Therapy in the Medical ICU
13	van Ruler	Journal of the American Medical Association	2007	Comparison of On-Demand vs Planned Relaparotomy Strategy in Patients With Severe Peritonitis: A Randomized Trial
14	Krueger	New England Journal of Medicine	2007	A Human Interleukin-12/23 Monoclonal Antibody for the Treatment of Psoriasis
15	Sandler	New England Journal of Medicine	2006	Paclitaxel-Carboplatin Alone or with Bevacizumab for Non-Small-Cell Lung Cancer
16	Dorsey	Journal of the American Medical Association	2007	Combination Therapy for Uncomplicated Falciparum Malaria in Ugandan Children: A Randomized Trial
17	Nissen	New England Journal of Medicine	2006	Effect of ACAT Inhibition on the Progression of Coronary Atherosclerosis
18	Taccone	Journal of the American Medical Association	2009	Prone positioning in Patients with Moderate and Severe Respiratory Distress Syndrome: A Randomized Controlled Trial
19	Dobscha	Journal of the American Medical Association	2009	Collaborative Care for Chronic Pain in Primary Care: A Cluster Randomized Trial
20	Perondi	New England Journal of Medicine	2009	A Comparison of High-Dose and Standard-Dose Epinephrine in Children with Cardiac Arrest

21	American Lung Association Asthma Clinical Research Centers	New England Journal of Medicine	2009	Efficacy of Esomeprazole for Treatment of Poorly Controlled Asthma
22	Montalescot	Journal of the American Medical Association	2009	Immediate vs Delayed Intervention for Acute Coronary Syndromes: A Randomized Clinical Trial
23	Kawamori	The Lancet	2009	Voglibose for prevention of type 2 diabetes mellitus: a randomised, double-blind trial in Japanese individuals with impaired glucose tolerance
24	Lennox	The Lancet	2009	Safety and efficacy of raltegravir-based versus efavirenz-based combination therapy in treatment-naïve patients with HIV-1 infection: a multicentre, double-blind randomised controlled trial
25	Albert	Journal of the American Medical Association	2001	Effect of Statin Therapy on C-Reactive Protein Levels: The Pravastatin Inflammation/CRP Evaluation (PRINCE): A Randomized Trial and Cohort Study
26	de Smet	New England Journal of Medicine	2009	Decontamination of the Digestive Track and Oropharynx in ICU Patients
27	Karunajeewa	New England Journal of Medicine	2008	A Trial of Combination Antimalarial Therapy in Children from Papua New Guinea
28	Barwell	The Lancet	2009	Comparison of surgery and compression with compression alone in chronic venous ulceration (ESCHAR study): randomised controlled trial

Appendix 2: The grid

Section: Selection Bias		Reported adequately?		Low risk of bias?	
	Issue	Judgment	Description	Judgment	Description
A.	□ **Randomisation generation and implementation** ○ Do the authors provide sufficient information that the reader can assess the methods used to generate the random allocation sequence and the likelihood of bias in treatment allocation? ○ Does the paper explain who generated the allocation sequence, who enrolled participants and who assigned participants to the trial group?	Yes No	Quote: Comment:	Yes No / Unclear	Quote: Comment:
	□ **Flow of participants -** does the paper state how many participants: ○ Were assessed for eligibility ○ Were eligible ○ Were enrolled ○ Were excluded ○ Were randomised to each intervention? ○ Are these numbers given in a clear, easily interpretable manner?	Yes No	Quote: Comment:	Yes No / Unclear	Quote: Comment:
	□ **Baseline demographics -** are the study groups compared at the baseline for important demographic and clinical characteristics, allowing the reader to assess how comparable they are?	Yes No	Quote: Comment:	Yes No / Unclear	Quote: Comment:

Section: Performance Bias		Reported adequately?		Low risk of bias?	
	Issue	Judgment	Description	Judgment	Description
B.	□ **Blinding and data collection –** participants are ideally blinded to their allocation status. Are the participants in the trial blinded? If participants are not blinded, are the study endpoints objective and collected by someone unlikely to influence the response differentially? (e.g. not data from self-reporting or someone affiliated with the intervention) If not, does the paper discuss the resultant risk of bias and what is done to control for it?	Yes No	Quote: Comment:	Yes No / Unclear	Quote: Comment:
	□ **Blinding and participant conduct** – again, participants are ideally blinded to their allocation status. Does the paper mention whether blinding recipients was possible and, if so, considered? If not, does it discuss the potential problems from participants seeking care differentially as a result of being aware of their treatment allocation and whether these problems are likely to have occurred?	Yes No	Quote: Comment:	Yes No / Unclear	Quote: Comment:

Section: Detection Bias		Reported adequately?		Low risk of bias?	
	Issue	Judgment	Description	Judgment	Description
C.	□ **Data collection -** does the paper state: ○ How the data is collected ○ Who is collecting the data ○ What relationship, if any, the data collectors have to the intervention? ○ Does the paper mention whether blinding data collectors was possible and, if so, considered?	Yes No	Quote: Comment:	Yes No / Unclear	Quote: Comment:

Section: Attrition Bias		Reported adequately?		Low risk of bias?	
	Issue	Judgment	Description	Judgment	Description
D.	□ **Flow of participants -** does the paper state how many participants: ○ Received each intervention ○ Did not receive each intervention ○ Were followed up ○ Were lost to follow up ○ Were included for analysis ○ Were excluded from the analysis by the investigators?	Yes No	Quote: Comment:	Yes No / Unclear	Quote: Comment:
	□ **Number of participants/intention to treat -** does the paper give the number of participants in each group included in the analysis, and whether this analysis is according to the "Intention to Treat" principle? If not, is there evidence that the principle was followed?	Yes No	Quote: Comment:	Yes No / Unclear	Quote: Comment:

Section: Reporting Bias		Reported adequately?		Low risk of bias?	
	Issue	Judgment	Description	Judgment	Description
E.	□ **Pre-specified protocol and analysis plan** - does the paper have a pre-specified protocol and analysis plan for conduct and evaluation of the trial?	Yes No	Quote: Comment:	Yes No / Unclear	Quote: Comment:
	□ **Outcomes and summary of results** ○ Are all presented outcomes defined as primary, secondary or exploratory? ○ Are the results presented for all planned primary and secondary endpoints? ○ Are the results presented in an intuitive manner, including the summary of each outcome and the measured effect size with a confidence interval?	Yes No	Quote: Comment:	Yes No / Unclear	Quote: Comment:

Section: Reporting Bias (cont'd)		Reported adequately?		Low risk of bias?	
	Issue	Judgment	Description	Judgment	Description
E.	□ **Ancillary analyses –** do the authors present or offer a link to an appendix listing the exploratory analyses performed but not presented in the paper?	Yes No	Quote: Comment:	Yes No / Unclear	Quote: Comment:
	□ **Interpretation -** does the interpretation of the results: o Offer a synopsis of the findings o Provide a consideration of possible mechanisms and explanations o Offer comparison with relevant findings from other studies and discuss the results of the trial in the context of existing evidence, evidence which is not limited to evidence that supports the results of the current trial o Discuss limitations of the present study o Exercise special care when evaluating the results of a trial with multiple comparisons (e.g. multiple endpoints or subgroup analyses)?	Yes No	Quote: Comment:	Yes No / Unclear	Quote: Comment:

Section: Sample Size		Reported adequately?		Low risk of imprecision?	
	Issue	Judgment	Description	Judgment	Description
F.	□ **Sample size -** do the authors indicate whether they conduct a sample size calculation and if so, how?	Yes No	Quote: Comment:	Yes No / Unclear	Quote: Comment: