RANDOMIZATION IN MEDICINE AND ECONOMICS

David Teira (UNED) & Julian Reiss (EUR)

[1st (rough) draft]

1.

In this paper we want to bridge the gap between philosophy of economics and philosophy of medicine and discuss the methodological virtues of randomized experiments in both fields. Randomized clinical trials constitute one of the pillars of contemporary medicine, and philosophers have been assessing their epistemic scope throughout the last three decades. The current philosophical consensus seems to be that RCTs do not provide an absolutely superior form of evidence, despite their popularity among clinicians and their prominent role in our pharmaceutical regulatory agencies. In the last decade there has been an explosion of interest on randomized field experiments in development economics. Philosophers have not paid so much attention to these latter (with the exception of Nancy Cartwright's 2010 PSA lecture), even if they often promise to bring about a revolution in our methods to fight poverty. At stake both in medicine and economics is the possibility of an evidence-based policy (be it for pharmaceutical regulation or for development).

In this paper, we present first an introduction to clinical trials in medicine, discussing the contribution of randomization to causal analysis in RCTs. Taking sides here with Cartwright and Worrall, we argue that the validity of randomization cannot be established *ex ante*, or by purely statistical considerations, but requires expert judgment. In the third section of the paper, we analyze the actual success of RCTs in regulatory drug testing. We show that the reliability of our regulatory agencies does not depend exclusively on RCTs (as their critics sometimes think), but on a system of checks and balances, which includes extensive post-marketing surveillance. This system relies on expert judgment and it is crucial for it being credible that the experts appear impartial. Randomization provides a good warrant of impartiality, preventing selection biases.

However, as we will see in detail in section 4, by itself it is not enough to secure the actual impartiality of RCTs: the financial stakes at each regulatory trial are high enough to elicit conflicts of interests at each stage of their design and implementation. Neither randomization nor any other single tool can cope with all the potential sources of bias in

a regulatory experiment. But, against some critics, we argue that this is a reason to keep randomization in medical trials, rather than dispensing with it: without randomization a drug test is more vulnerable to bias than without it.

However, this might not be the case of randomization is development economics. As we will see in the fifth and final section of this paper, allocating treatments at random can interfere with the decisions of the potential participants, creating a Hawthorne effect. They will not behave in the experiment as they would have without it. Randomization does not warrant here either the impartiality of the experiment, because it creates an incentive to manipulate the results according to the preferences of the participants (once they realize in which group they are). Inviting third parties to run the experiment (international organizations) is not such a good solution as the conduct of drug trials by regulatory agencies in medicine.

2. RANDOMIZATION AND CAUSALITY IN RCTS

Stuart Pocock (1983) defined clinical trials as planned experiments, involving patients with a given medical condition, which are designed to elucidate the most appropriate treatment for future cases. The canonical example of experiments of this sort is the drug trial, which is usually divided into four phases[1]. Phase I focuses on finding the appropriate dosage in a small group of healthy subjects (20-80); thus such trials examine the toxicity and other pharmacological properties of the drug. In phase II, between 100 and 200 patients are closely monitored to verify the treatment effects. If the results are positive, a third phase, involving a substantial number of patients, begins, in which the drug is compared to the standard treatment or placebo. If the new drug constitutes an improvement over the existing therapies and the pharmaceutical authorities approve its commercial use, phase IV trials are begun, wherein adverse effects are monitored and morbidity and mortality studies are undertaken.

This paper focuses on phase III drug trials. The standard experimental design for these trials currently involves a randomized allocation of treatments to patients. Hence the acronym RCTs, standing for randomized clinical (or sometimes controlled) trials[2].

---

[1] Clinical trials can be set to analyse many different types of treatment: not only drugs, but also medical devices, surgery, alternative medicine therapies, etc. The characteristics of these types of trials are quite different; so, for the sake of simplicity, I will only deal here with drug testing.

[2] In this thesis, for the sake of simplicity, "RCTs" will refer to standard frequentist trials. Notice though that randomization may well feature in the design of a Bayesian clinical trial.

The statistical methodology for planning and interpreting the results of RCTs is grounded in the principles established by Ronald Fisher, Jerzy Neyman and Egon Pearson in the 1920s and 1930s. A hypothesis is made about the value of a given parameter (e.g., the survival rate) in a population of eligible patients taking part in the trial. The hypothesis is tested against an alternative hypothesis; this requires administering the drug and the control treatment to two groups of patients. Once the end point for the evaluation of the treatment is reached, the interpretation of the collected data determines whether or not we should accept our hypothesis about the effectiveness of the drug, assigning a certain probability to this judgment. This statistical methodology is based on a specific view of probability, frequentism, according to which probabilities are (finite or infinite) relative frequencies of empirical events: the treatment effects in a given experimental setting. The first clinical trial planned and performed following a frequentist standard was the test of an anti-tuberculosis treatment, streptomycin. It was conducted in Britain and published in 1948. Over the following decades, RCTs would be adopted as a testing standard by the international medical community and by pharmaceutical regulatory agencies all over the world. Today, RCTs constitute the mainstream approach to drug testing and, through evidence-based medicine, they even ground standards of medical care.

Running a phase III clinical trial is a complex task, which goes far beyond its statistical underpinnings. The credibility (and feasibility) of a trial is conditional on a complete preplanning of every aspect of the experiment. This plan is formally stated in the study protocol. The following items should feature in the protocol, according again to Pocock (1983, p. 30):

Background and general aims	Patient consent
Specific objectives	Required size of study
Patient selection criteria	Monitoring of trial progress
Treatment schedules	Forms and data handling
Methods of patient evaluation	Protocol deviations
Trial design	Plans for statistical analysis
Registration and randomization of patients	Administrative responsibilities

The aim of a trial is to test a hypothesis about the comparative efficacy of an experimental treatment (be it with the standard alternative or a placebo). Leaving aside for a moment the statistical design of the test, first it is necessary to define which patients are eligible for the study (e.g., they should be representative of the disease under investigation); how to create an experimental group and a control group; how to administer treatment to each of them; and what the end-points for the evaluation of their responses are. During the course of the trial, an *interim analysis* is usually performed in order to monitor the accumulating results, since reaching the number of patients specified in the design may take months or years and because the information gleaned from such an interim analysis may in fact warrant some action such as terminating the trial early. Once the trial is completed, the hypothesis about the comparative efficacy of the treatment will be either accepted or rejected and the results published. Depending on the disease and the planned sample size, this may add several years to the time taken up by the two previous trial phases. Thus the development of a new drug may well take a decade before it is approved for public use by the pharmaceutical regulatory agency[3].

Once a patient is deemed eligible (according to the trial's protocol) and recruited, the informed consent form signed and the log sheet with her identification details filled out, the treatment is assigned at random. Depending on the arrangement of the trial (number of treatments, whether or not it is double blinded, whether or not it is multi-centre), randomization may be implemented in different ways. The general principle is that each patient should have an equal probability of receiving either treatment. If it is convenient to control the allocation of treatments according to patient characteristics, in order to prevent imbalances, randomization can be *stratified*. For most practitioners, randomization accounts for the ability of RCTs to detect causal mechanisms. The justification often can be traced back to Fisher's famous *tea tasting* experiment. In clinical trials, randomization would allow control over unknown prognostic factors, since, over time, their effect would be distributed in balance between the two groups.

However, the statistical rationale of the procedure is slightly more complex. Let us informally sketch Fisher's original argument for randomization (as reconstructed in (Ba-

[3] For an updated account of the practical arrangements involved in a trial, including compliance with the current regulatory constraints, see Hackshaw 2009, pp. 157-201. The book provides a concise overview of every dimension of a clinical trial today. For a thorough overview, see Piantadosi 2005.

su, 1980)). In order to study the response differences between the two treatments in trial patients, we need a test statistic with a known distribution: for instance, $T=\Sigma d_i$, where d_i is the response difference. Assuming the hypothesis that there is no difference between treatments, suppose we observe a positive difference d_i between treatments in a given pair of patients who received them at random. Assuming that our hypothesis is true, this difference must have been caused by a nuisance factor. If we kept this factor (and all other factors) constant and repeated the experiment, the absolute value of $|d_i|$ would be the same with the same sign, if the treatments were identically allocated; it will be reversed if the allocation had been different. The sample space of T will be the set of 2^n vectors $R = \{(\pm d_1, \pm d_2, \ldots, \pm d_n)\}$

Randomization warrants that all these vectors will have an equal probability. If d_i is positive for all i, we will observe another n response differences d'_i equal or bigger than d_i if and only if $d'_i = d_i$. The probability of observing this response is $(½)^n$, the significance level of the observed differences, as we will see below. This probability was, for Fisher, the frequency of observing this difference in an infinite series of repetitions of the experiment. And we will need it in order to calculate how exceptional the results of our experiment have been (e.g., *via* p-values).

In order to detect causal mechanisms through standard RCTs we need to assume thus a frequentist conception of probability and interpret randomization in its light. In addition, we need to adopt a probabilistic view of causality which can make sense of RCTs. Let us illustrate it with Nancy Cartwright's analysis. According to Cartwright, RCTs are one method, among others, for warranting causal claims. It is a *clinching* method that proceeds in a deductive fashion: if its assumptions are met and the necessary evidence is provided, we can safely affirm the concluding causal claim. However, the premises are of a very restrictive form and, therefore, the conclusions are narrow.

RCTs test causal claims about the safety and efficacy of drugs in a given population following Mill's *method of difference*. Given a selected outcome (O), we study the probability of the difference between outcomes with and without the treatment intervention (T) in two groups of patients drawn from a population φ. In these two groups, every causally relevant factor other than T are equally distributed. Hence, the observed difference in O would be an effect of T. We need a number of assumptions to make sense of this causal claim.

The first assumption is a *causal fixing condition* (Cartwright and Munro 2010, p. 261): the probability of an effect is fixed by the values taken by a full set of its causes. Cartwright adopts a version of Suppes' probabilistic theory of causality, that generally states that for an event-type T temporally earlier than event-type O in a population φ,

$$T \text{ causes } O \text{ in } \varphi \text{ iff } P(O/T\&K_i) > P(O/\neg T\&K_i)$$

$$\text{for some subpopulation } K_i, \text{ with } P(K_i) > 0$$

We are assuming in addition that the individuals in our sample are all governed by the same causal structure[4], described by a probability distribution P. According to Cartwright, "P is defined over an event space $\{O,T,K_1,K_2,\ldots,K_n\}$, where each K_i is a state description over 'all other' causes of O except T"[5]. Conditioning on these potential *confounding factors*, we can attribute the remaining difference between $P(O/T\&K_i)$ and $P(O/\neg T\&K_i)$ to the causal link between O and T.

For Cartwright, randomization would be just a means of controlling the probabilistic dependences in the experimental and the control group , making sure that they are the same in both wings except for C. In an ideal RCT, claims Cartwright (2007, p. 15), the K_i are distributed identically between the treatment and control groups. Hence any difference in outcome between groups can be causally attributed to T in at least one K_i relative to the causal structure (CS) described by P. This is the conclusion that RCTs can clinch. However, according to Cartwright, we need additional assumptions if we want to generalize this conclusion to some target population φ.

If we want to affirm, for instance, that T causes O in at least some members of Θ, Cartwright (2007, p. 17) argues, we need assumptions of this kind:

- At least one of the subpopulations (with its particular fixed arrangement of 'other' causal factors) in which T causes O in φ is a subpopulation of Θ.
- The causal structure and the probability measure is the same in that subpopulation of Θ as it is in that subpopulation of φ.

The main warrant for these assumptions in RCTs provide comes thus from randomization However, throughout the last three decades, philosophers of science (namely Peter Urbach and John Worrall) have put in question this view of randomization. Let us list the main objections:

[4] In Cartwright 2007, p. 16, this causal structures are defined as "the network of causal pathways by which O can be produced with their related strengths of efficacy".

[5] K_i are "maximally causally homogeneous subpopulations" of φ: see Cartwright 2010, pp. 61-63 for a discussion.

Objection #1: which population?

In a clinical trial there is no real random sampling of patients, since the population random samples should be drawn from remains usually undefined: there is no reference population, just criteria of patient eligibility in the trial protocol. Generalizing from the individuals entered into the study to any broader group of people seems ungrounded (Urbach 1993).

Objection #2: significant events may not be that rare

A positive result in a significance test is interpreted as an index that H_0 is false. Were it true, such result would be an "exceptionally rare chance". It would be exceptional because a randomized allocation of treatments would ideally exclude any alternative explanation: uncontrolled factors would be evenly distributed between groups in a series of random allocations. However, it would not be "exceptionally rare" that the treatment was effective in the case where it had been allocated to the healthier patients alone, to those with best prognoses or to any group of patients that for whatever reason could differentially benefit from the treatment.

Colin Howson, among others, has argued that randomization as such does not guarantee that the occurrence of such unbalanced allocation *in a particular trial* is rare: it may be produced by uncontrolled factors. As Worrall (2007, pp. 1000-01) puts it, "randomization does not free us from having to think about alternative explanations for particular trial outcomes and from assessing the plausibility of these in the light of 'background knowledge'". This further assessment cannot be formally incorporated, as it should be, into the methodology of significance testing. Hence, we cannot ground our conclusions on this methodology alone.

Objection #3: post randomization selection

By sheer chance, a random allocation may yield an unbalanced distribution of the two test groups, i.e., the test groups may differ substantially in their relevant prognostic factors (these are called *baseline imbalances*). This difference may bias the comparison between treatments and spoil the experiment. If one such distribution is observed, the customary solution is to randomize again seeking a more balanced allocation. However, argues Urbach (1985), the methodology of significance testing forbids any choice between random allocations: if they are adequately generated, any allocation should be as good as any other. Hypotheses should be accepted or rejected on the basis of the expe-

riment alone, without incorporating our personal assessment of the generated data (justified though it may be).

It is usually assumed that with a high number of enrolled patients, it is very unlikely that randomization generates unbalanced groups. Urbach argues that we cannot quantify this probability and much less discard it. At best, a clinical trial provides an estimation of the efficacy of a treatment, but there is no direct connection between this result and the balance of the two groups. The conclusions of the trial can be spoiled by the following two objections.

Objection #4: unknown nuisance variables after randomization

Even after randomizing, uncontrolled factors may differentially influence the performance of a treatment in one of the groups. Further randomizations at each step in the administration of the treatment (e.g., which nurse should administer it today?) may avoid such interferences, but this is quite an impractical solution. Declaring such disturbances negligible, as many experimenters do, lacks any internal justification in the statistical methodology assumed (Urbach 1985, Worrall 2007).

Objection #5: known nuisance variables

It has been argued that randomization can at least solve the problem created by known perturbing factors that are difficult to control for. These could be at least randomized out. Following Levi (1982), Urbach (1985, p. 267) argues that since we know of no phenomena correlated to these confounding factors, "there is no reason to think that they would balance out more effectively between groups by using a physical randomizing device rather than employing any other method".

In sum, randomization per se does not guarantee the internal validity of an RCT: the objections show that we need the judgment of an expert to judge whether randomization has been properly conducted. As Nancy Cartwright puts it:

> Without expert judgment, however, the claims that the requisite assumptions for the RCT to be internally valid are met depend on fallible mechanical procedures. Expert judgments are naturally fallible too, but to rely on mechanics without experts to watch for where failures occur makes the entire proceeding unnecessarily dicey. (Cartwright 2007, p. 19)

2. CAUSALITY AND THE REGULATION OF THE PHARMACEUTICAL MARKETPLACE

As we said in the introduction, RCTs provide the standard of proof for regulatory agencies and we may wonder to what extent the objections examined in the section above impinge on their practice. We will assess the external validity of RCTs in a regulatory context, showing that despite their flaws, RCTs are still considered a reliable regulatory tool by the American public. We claim that randomization provides a warrant of impartiality that crucially contributes to this consideration, precisely because it constraints expert judgment (which, as we will see, is necessary to assess the RCT results). In order to make our case, let us briefly present first the regulatory role of RCTs in the U.S. pharmaceutical markets.

The 1950s saw a boom in industrial drug production (some were *wonder drugs*, e.g. antibiotics, but many were just combinations of already available compounds) and, simultaneously, in pharmaceutical advertising that caused much confusion among practitioners about the therapeutic merit of each product (Marks 2000, p. 346)[6]. For a minority of therapeutic reformers, RCTs with a strict research protocol provided the information about drugs that "sleazy advertising" was trying to disguise with "badly scissored quotes", "pharmaceutical numbers racket", "detail men" visits and so forth (Lasagna 1959, p. 460-461)[7]. Hence advising physicians to prescribe on the basis of RCT-based evidence was perceived as necessary. Moreover, the increasing commercial pressure on pharmaceutical research (as RCTs became part of the advertising engine) made it necessary to enforce the observance of a rigorous trial protocol[8].

In the 1960s, defenders of a more severe pharmaceutical regulation in Congress found in the thalidomide scandal the lever needed to pass what would become the 1962 Drug efficacy amendment to the Food, Drug and Cosmetics Act. It required from the applicant "adequate and well-controlled clinical studies" for proof of efficacy and safety. The FDA developed this indication in further detail: The "Form FD-356", published in the

---

[6] See, e.g., the editorial pleads in the *Journal of the American Medical Association* 165 (1957), p. 688 and the *New England Journal of Medicine* 258 (1958), p.145.

[7] For a quick description of the commercial distortions of pharmaceutical research in the United States and the role that clinical trials could play to regulate it see Dowling 1957 and Sheps 1961.

[8] "To get an approximate idea of the current situation in all specialties, the author reviewed the summaries of original investigations which appeared in the *1959-60 Year Book of Drug Therapy*. Out of 394 summaries which gave adequate information about the plan of the study, 225 (57 per cent) related to reports without any explicit comparison with the results of another treatment in similar patients. Every section of the *Year Book* contained reports which ignored the hard fact of unpredictability and simply described the course followed by patients while they were being given a drug. All of this evidence suggests that a great part of the so-called clinical evaluations of new drugs is unscientific, lacks adequate provisions to eliminate bias, and cannot be objectively judged." (Sheps 1961, p. 651)

Federal Register on February 14, 1963, established (Carpenter & Moore 2007, pp. 355-356):

> An application may be incomplete or may be refused unless it includes substantial evidence consisting of adequate and well-controlled investigations, by experts qualified by scientific training and experience to evaluate the effectiveness of the drug involved, on the basis of which it could fairly and responsibly be concluded that the drug will have the effect it purports or is represented to have under the conditions of use prescribed.

Although the definition of a well controlled investigation would not be clarified until 1969, when it was formally quantified as two well-controlled clinical trials plus one previous or posterior confirmatory trial (Ceccoli 2004, p. 181), Carpenter and Moore claim that this set of regulations created the modern clinical trial industry[9]. In the coming three decades, pharmaceutical funding would propel the conduct of RCTs in the United States (and elsewhere).

How can we measure the performance of regulatory RCTs? Have they adequately assessed the efficacy and safety of the compounds introduced in our pharmaceutical markets? More precisely, the purported causal effects detected in trials have they been adequately generalized to the corresponding patient populations? It is worth recalling that the Food and Drug Administration (FDA), like other regulatory agencies, does not take the external validity of the RCTs leading to the approval of a new substance for granted. There is a fourth phase in clinical trials, post-market surveillance in which the FDA collects adverse event reports from various sources and conducts epidemiological studies to assess their relevance. In other words, the FDA keeps track of the validity of the results of their trials in the general population.

However, the authority and resources of the FDA at this stage are disproportionately smaller than at any previous point in the approval process. And the assignment is big: apart from monitoring adverse reports, the agency has to consider as well issues in labeling, advertising or the inspection of production and storage facilities, to name just a few. Hence, we cannot take the market withdrawal of a drug as the best indicator of the

[9] For analysis of several paradigmatic trials after the 1963 amendment, see Marks 1997, pp. 164-229 and Meldrum 1994, pp. 310-372.

lack of external validity of the antecedent trials[10]. As Dan Carpenter (2010, ch. 9) has argued, the negotiation of each withdrawal depended on a number of circumstances outside and inside the agency.

Among these latter circumstances, we should mention the conflict between the standards of proof employed in the third and fourth phase of the trial. A group of pharmacologists who came to dominate the FDA in the 1960s imposed the RCT as the main evidentiary standard. Still in our days they are the leading voice within the agency and epidemiologists continue playing a secondary role (Carpenter 2010, p. 612). This creates a sort of pharmacologists' confirmation bias at the agency. These pharmacologists, we may suspect, refuse to revise their approvals in the light of additional epidemiological evidence to the contrary[11].

A more reliable indication of minor or major failures at phase 3 trials are changes in drug labeling. According to Dan Carpenter (2010, p. 212), the FDA has relied on these changes as a cheap regulatory strategy, given the available resources, as compared with pursuing withdrawal or a change in advertising and prescribing practices – which in the United States are only lightly regulated. As long as the label describes potential safety threats, the FDA can claim that the consumer has been warned. Each label change requires an application for approval, which creates a data record. Dan Carpenter has compiled it in the following table, where it is compared to other product changes for the same periods:

[10] For an analysis of the complex circumstances of a withdrawal, see Abraham and Davis 2009 and Abraham 2010.

[11] Dan Carpenter (2010, pp. 623-626) has found that for every year increase in the average staff tenure at the approving division of the FDA the drug is 6.9 percent less likely to experience a black-box label warning. Carpenter suggests we may conjecture that with longer tenure there will be growing concern in the FDA officers about their reputation; therefore, they will resist the admission of failure in their judgment of a drug. For parallel evidence regarding judicial convictions see Tavris and Aronson 2007, chapter 5.

Drug Changes Requiring a Supplemental NDA, 1970–2006

	1970–1974	1975–1979	1980–1984	1985–1989	1990–1994	1995–1999	2000–
Chemistry Revisions	2	376	3,710	7,728	5,664	8,520	258
Manufacturing Revisions	0	492	910	1,045	1,063	2,229	1,936
Package Changes	38	465	757	733	573	847	994
New or Modified Indications	3	6	7	76	121	273	294
Control Supplements	242	2,516	3,710	2,138	1,902	2,885	4,357
Labeling Revisions (SLR)	529	1,968	2,005	2,360	1,909	2,341	4,472
Other Label Changes	0	0	168	1,998	3,588	2,923	1,672

Source: FDA, Drugs@FDA database.

Table 2 (Carpenter 2010, p. 613)

Carpenter (2010, p. 623) summarizes it as follows: from 1980 to 2000 the average new molecular entity received five labeling revisions after approval, about one for every three years of marketing after approval. Only one in four drugs had no labeling revisions at all. Even if the data come mostly from post-Bayh-Dole RCTs, it seems exaggerated to attribute all this revisions to the biases introduced by the trial sponsor –though this is something that only a statistical analysis can conclude.

The data are obviously too rough to decide what went wrong, if anything, in the phase 3 RCTs: the sample might have been too homogeneous as compared to the patients that finally used the therapy, as Cartwright suggested, or the trial too brief to detect adverse effects (e.g., toxicity or cardiovascular outcomes). If we take external validity in an equally rough sense, Carpenter's data suggest that Cartwright points out correctly the limitations of causal inference in RCTs. In the end, we need experts to decide whether the data grant the regulatory approval and their judgment may fail.

However, we do not think that the FDA or any other regulatory agency ever were so ambitious as to assume that the external validity of phase III RCTs could be taken for granted without further checks. The establishment of a monitoring system on approved drugs shows that the FDA acknowledged the fallibility of phase III trials and possibility

of making incorrect decisions about drug approvals. This is a first crucial point we should bear in mind: *the reliability of clinical trials for regulatory purposes depends not only on RCTs, but on an entire institutional system of checks and balances*. The first two phases of a trial provide background knowledge on which the RCTs rely and their conclusions should be later validated by extensive monitoring.

Imperfect as it may seem, the U.S. public has considered the FDA a reliable regulator: it may fail, *but not in the interest of a private party* (Carpenter 2010). We contend that even if it is an imperfect means to detect causal connections, randomization has played a crucial role in granting the independence and credibility of regulatory agencies, if only for one reason: it prevents *selection biases*. Medical experiments have indeed straightforward economic consequences: the inventor of successful therapies may become rich at the expenses of those who have to pay it, be it the consumers or their health insurance. These financial incentives are prone to generate biases and may explain as well why the results of an experiment are contested if it does not accord with the expectation of one of the parties. Randomization prevents investigators from assigning (consciously or unconsciously) patients with, say, a given prognosis to any one of the treatments. For instance, an investigator might allocate the experimental treatment to the healthier patients, if she wants the trial to be positive, or to the patients with a worse prognosis, if she thinks they will benefit more.

From a purely causal standpoint, randomization is just a means of controlling the probabilistic dependences arising from biases in the experimental and the control group, making sure that they are the same in both wings except for the treatment. For the pharmaceutical consumer, randomization appears as a warrant of the impartiality of the test: nobody will allocate treatments for his own particular benefit. Even if randomization requires expert judgment, it also constrains it in a way that seems desirable from a regulatory viewpoint.

Between 1900 and 1950 pure expert clinical judgment was the main approach in the assessment of the properties of pharmaceutical compounds, both in Britain and the United States. An experienced clinician would administer the drug to a series of patients he considers more apt to benefit from it. His conclusions would be presented as a case report, with the details of each patient's reaction to the treatment. The alternatives were first laboratory experiments and then controlled clinical trials (from which RCTs would later emerge). The former would proceed either in vitro or in vivo (on animals and pa-

tients): considered superior by clinicians with a scientific background, its scope was usually restricted to safety considerations. It soon gave way to comparative trials, in which two treatments were alternated on the same patient or administered in two groups of patients (simultaneously or not). The arrangements to secure the comparability of the two treatments were the controls and they adopted different forms: eligibility criteria, alternation and then randomization in the allocation of treatments, uniformity in their administration and blinding were the most prominent. They were not used necessarily all at once. Statistical reports from controlled trials conveyed their results with different degrees of sophistication. Significance testing features only occasionally before 1950.

The regulatory authorities in Britain and the United States arranged official drug testing depending on the standards adopted by the research community within their respective medical professions. In both cases, and throughout the 20th century, the regulators were concerned about impartiality, understood here as *independence from the financial interests of the pharmaceutical industry*. Tests sponsored by manufacturers for advertising purposes were considered suspicious by the consumers in both countries and this prompted, in different ways, the development of public pharmaceutical agencies to conduct or supervise the tests. However, most clinical researchers considered themselves impervious to biases from non-financial sources and impartial enough to conduct clinical tests without selection-proof mechanisms. Until the 1960s, regulatory decisions were fundamentally based on expert judgment of this sort. Expert judgment came only to be discredited in the United States when the pharmaceutical industry became big enough to advertise at such a massive scale that the judgment of individual clinicians ceased to be considered reliable at the regulatory agencies.

However, as Iain Chalmers and Harry Marks have argued, the inferential power of RCTs and their statistical foundations were not the primary reason to adopt them: randomization, blinding and significance testing were *impersonal* rules to allocate treatments and interpret the results of a trial and, therefore, warrants of impartiality against the interests of the pharmaceutical industry. Rather than dispensing with expert judgment, these debiasing rules constrained it in order to make it credible to the consumer. A clinical trial might be inherently uncertain from a causal point of view and yet we may well accept its results if we are granted that nobody exploits this uncertainty for his own private benefit.

To sum up, randomization might be defended, in the context of a regulation, as a warrant of impartiality, as a constraint on the judgment of the experts, even if judgment is nonetheless necessary to establish the causal scope of RCTs.

3. THE ACTUAL IMPARTIALITY OF RCTS

We may now wonder to what extent the warrant of impartiality is enough to secure the fairness of a trial. As of today, the answer seems not very promising: the conflicts of interest pervading biomedical research seem to overflow whatever warrant set against them in RCTs. Again, we need to put this claim in context.

In 1980 the Bayh-Dole Act gave the pharmaceutical industry the chance to intervene more deeply in the organization of biomedical research, including the running of RCTs[12]. By 2005 only 25% of all pharmaceutical research was conducted in academic medical centers (the figure was 80% before 1990: Fisher 2009, p. 4). The industry was spending around $15 billion in clinical trials, nearly three times as much as the entire public budget in the United States (Fisher 2009, p. 5). The money was allocated to an array of companies involved in the conduct of trials. Patients participating in the trial are treated through the so-called *investigative sites*, all sort of facilities ranging from university hospitals to small private practices. Sometimes they are centrally hired and coordinated by *Site Management Organizations*. The pharmaceutical industry usually deals with them through *Contract Research Organizations* that hire the necessary sites for each trial and monitor its conduct. The trials are also externally monitored by professional Institutional Review Boards, checking their scientific and ethical standards on demand for a fee. There are also auxiliary companies dealing with the advertising and recruitment of patients, among other aspects of the trial. The main justification of this research market is the possibility of speeding each stage of the trial in order to get as quickly as possible to the consumers and make the most of each patent[13].

In this context, an increasing concern has emerged about the conflicts of interest that biomedical researchers are facing. These conflicts are informally understood as potential sources of bias: in a trial the duties of the clinician (e.g., patients' care) or scientist

[12] For a complete survey and discussion of the effects of the Bayh-Dole Act on North-American science, see Krimsky 2003.

[13] New compounds are patented at the start of clinical research, not after FDA approval: the three phases of a trial can consume up to nine years and less than a third of the trials are successful. Thanks to this for-profit organization of research, companies were filing applications to the FDA in 4.9 years on average between 1999 and 2003, whereas in the period 1994-1998 it had taken 5.4 years (Fisher 2003, pp. 12-13). The FDA is speeding their review process as well after introducing fees for the filing companies to subsidize the process.

(e.g., truth finding) can be influenced by their personal interests at stake in the research, be these financial (e.g., as stockholders) or social (e.g, status)[14]. Little is known about the precise belief-forming mechanisms at operation in these conflicts. Nonetheless there is increasing evidence, for instance, about the effects of gifts on the prescriptions of practicing physicians[15]. We can presume that their colleagues conducting trials may be no less sensitive to the industry attentions.

Starting in the 1980s, a growing number of quantitative studies were conducted in order to estimate the scope and impact of these conflicts of interests among researchers. In the early 2000s, the first systematic reviews aggregating data from all these studies were published. The two more influential of these meta-analyses probably still are Bekelman et al. 2003 and Lexchin et al. 2003. Financial interests are loosely defined encompassing such different relationships with the industry as sponsorship, consultancy, employment, technology transfer, new venture formation, gifts or personal funds. The outcome varies depending on the aim and of each type of study and the methodology implemented.

Bekelman et al. (2003, p. 455) considered the proportion of industry sponsored and non-industry-sponsored studies with a certain outcome or characteristic. They reviewed 8 papers that analyzed the association between the sponsor and the study outcome in 1140 studies (not all of them RCTs). Assuming that the industry prefers a positive outcome for the experimental treatment – since they are investing in its development – the 8 papers reviewed yielded a higher probability of achieving such positive conclusion under private than under public sponsorship[16].

Was the association between conclusions and sponsorship due to the methodological quality of the original studies? Five of the eight papers reviewed in Bekelman et al. (2003, p. 459) reported that the quality of the studies was comparable, whatever the sponsor. Leaving aside the meta-methodological discussion of whether methodological quality can be measured, it is interesting to notice that many of these quality scales for RCTs hinge on bias control. The Jadad scale (Jadad et al. 1996) used in several papers analyzed in Bekelman et al. (2003) is built on the following three questions:

[14] The standard definition of a conflict of interest in the biomedical literature is Denis Thompson's: "a set of conditions in which professional judgment concerning a primary interest (such as a patient's welfare or the validity of research) tends to be unduly influenced by a secondary interest (such as financial gain)'' (Thompson 1993). For an actual sample of these conflicts, see Murphy 2004, pp. 127-152.

[15] For a quick review, see Jain 2007.

[16] This is measured by the following odds ratio (p1/1- p1)/ (p2/1- p2), where p1 is the probability of a positive conclusion under private sponsorship and p2 under public sponsorship.

1. Was the study described as randomized (this includes the use of words such as randomly, random, and randomization)?

2. Was the study described as double blind?

3. Was there a description of withdrawals and dropouts?

In short, despite strict compliance with randomization and other debiasing rules that were regarded to warrant the fairness of RCTs only three decades ago, there seem to be pro-industry biases systematically interfering in the conduct of RCTs today. As Lexchin et al. (2003, p. 7) put it, "the results apply across a wide range of disease states, drugs, and drug classes, over at least two decades and regardless of the type of research being assessed". A simple look at table 2 will show that these three impartiality warrants cannot control all the potential sources of bias even at the stage of either designing a standard RCT. We owe this list to Lisa Bero and Drummond Rennie who adopt the physician's standpoint in the assessment of a drug study: it should provide "data on the relative effectiveness of a new drug: the clinical effectiveness, toxicity, convenience, and cost compared with available alternatives" (1996, p. 209). Effectiveness is defined, following the World Health Organization, as "the likelihood and extent of desired clinically relevant effects in patients with the specific indication" (*ibid.*).

Physicians interested in actual therapy care more about effectiveness than about mere efficacy, which usually covers whatever effect a drug may have independently of its therapeutic value. Despite their methodological quality, privately sponsored trials may be designed with a view to test for effects whose relevance for prescribing the drug is not accurately presented in the study. According to Bero and Rennie, a physician should consider such studies *biased*.[17]

[17] However, as some industry researchers suggest (Sacristan et al. 1997) biases may go in all directions: governments want to reduce their health budgets and they may prefer trials favoring the cheapest drug, which is often the control therapy. A "positive result" should thus indicate the result which coincides with the interests of the sponsor, rather than the one or another therapy. However, there is little evidence as of today regarding this second sort of bias.

I. The research question
    1. The research question is too narrow
II. Design of the study
    1. The study patients are not representative
    2. The allocation of treatments is not at random
    3. The choice of comparison groups favors the new drug
    4. No blinding of study subjects/researchers to treatment
    5. Use of fixed dosage
    6. Inappropriate choice of outcome measures
    7. Misleading data presentation and analyses
III. Conducting drug studies
    1. Failure to follow the established protocol
    2. Failure to keep records
    3. The submitted manuscript does not contain the data actually collected
IV. Drawing conclusions from the results of the study
    1. The conclusions do not relate to the research question
    2. The conclusions do not agree with results
V. Publication of drug studies
    1. The data are not published
    2. The funder controls the content of the publication
3. The funder bypasses the peer review process

Table 2: Potential Sources of Bias (Bero & Rennie 1996)

Leaving aside those items in the list that tend to appear only in trials of questionable quality (e.g., II.1, 2, 4, 7 and III), there are ways in which a researcher may raise his chances of getting positive results. For instance, the choice of dosage (II.5) and outcome measures (II.6). A researcher is free to choose a high dose of the experimental drug and any number of surrogate end points to assess its effect, whatever their clinical relevance. According to Bero and Rennie (1996, pp. 218-219), prescribing physicians should care instead for dose-response curves for the new drug compared to its competitors and actual health outcomes for each point in the curve. Researchers are equally free to chose the control treatment in a trial, be it another drug or a placebo (Bero and Rennie 1996, pp. 216-218). There are diseases (e.g., depression) for which the response to placebo is substantial and highly variable and a comparison with an already established drug will be probably favorable to the new treatment, but such trial will not deliver much information on relative effectiveness. Between 1945 and 1969, 100% of the trials were pla-

cebo controlled, whereas only 24% were between 1980 and 1989 (Bero and Rennie 1996, p. 218). The comparison with another drug gives also room for choice: it can be an alternative in the same drug class, in another drug class or even nondrug alternatives. In all cases the comparison is legitimate for research purposes, but prescribing physicians may find it more or less relevant for their clinical practice.

Costs are crucial for all parts, even if they are not explicitly included in the trial. On the side of the producer, it may happen that providing all the information demanded by the physician requires a more expensive trial. According to Bero and Rennie (1996, pp. 212-213), physicians want to know about relative treatment costs, a variable not always explicitly considered in the trial: depending on the dosage, side effects and available alternatives, a new drug may be more or less expensive to prescribe. Even if the study does not properly address all these concerns, the conclusions sometimes suggest the opposite. There is a growing industry of professional medical writers who are hired to produce more convincing papers, even if not always faithful to the study conducted.

Peer-review should detect and amend these conclusions in scientific journals, but it not always does and, moreover, there are many other medical publications (throwaway journals, symposia proceedings) that reach physicians without a referee control. We should also consider here the so-called publication bias all those negative results obtained in studies that never see the light (in print).

Strictly speaking, there might be no error in the trial: the background theory may be correct, the trial may be well designed and performed with proper caveats to check the assumptions and observational reports, and the results faithfully presented. This is why they are published and accepted as evidence by the regulatory agencies. Yet Bero and Drummond claim that from a clinical point of view, the published trial may be still biased, either because it answers clinically not very relevant questions or appears to answer relevant ones with no actual evidence to do it.

A simple solution are reader's guides providing checklists to verify the actual scope of a tria (e.g., Montori et al. 2004) l. But this new minority of medical reformers wants to debias the scientific journals themselves. Among the most popular measures proposed we should cite three: financial disclosure, reporting standards and trial registries. The first two rules should be implemented in the editorial policies in scientific journals: full disclosure of the financial relationships between researchers and the pharmaceutical industry should be a compulsory pre-requisite of publication; the trial should be reported

according to an agreed standard such as CONSORT. Finally, the pharmaceutical regulator should require that all trials are publicly registered to keep a record of the results obtained even if these are never published later.

All these are sensible and feasible policies, already in operation (to different degrees: see Ancker & Flanagin 2007). However, it has been argued that, even at their best, they will not completely correct the sort of biases denounced by Bero and Drummond. For instance, M. Doucet and S. Sismondo (2008) contend that in countries (like the USA) where 70% of the funding comes from the pharmaceutical industry, financial disclosure just reveals the obvious and, most often, it does not set any standard of comparison between potentially biased and unbiased trials. Reporting standards or trial registration policies may correct publication bias or distorted interpretations but, as Doucet and Sismondo put it, they do not set clinical relevance guidelines for trial design. Therefore, they cannot correct items II.3, 5, 6 in Bero and Drummond's list. As these latter suggest, we can only get such unbiased information if either it is legally required by the regulatory agencies in order to grant approval or these agencies conduct a study on their own to collect the missing data in each application.

Indeed, today many observers are advocating conduct and publication of second and third phase clinical trials in publicly funded institutions, either for a fee or directly "from the public purse"[18]: the results would become a public good available to all who request them. The industry would be free to organize their own trials, but for regulatory purposes only publicly financed trials would count for approval[19]. We think this is a sensible approach, even if the details of the proposal are still to be worked out. If expert judgment is both fallible and inevitable in the design and interpretation of trials, as the argument in section 1 and 2 suggests, we had better arrange an institutional setting in which the incentives for these experts do not collide with the public interest.

In sum, randomization may be an imperfect means to detect causal connections, requiring expert judgment to assess its results. It constrains the choices of the experts at the time at the time of allocating treatments, preventing selection biases. However, biases operate at different levels and randomization by itself is not enough to avoid them and guarantee the impartiality of a trial. Yet, this is not an argument for the dispensability of

[18] Prominent advocates of this approach are John Abraham (2010), Marcia Angell (2004), Sergio Sismondo (2008) and Boldrin and Levine (2008).
[19] There are alternative, but somewhat complementary approaches, such as the implementation of an adversarial system at the FDA defended by Justin Biddle (2007) or Julian Reiss.

randomization (e.g., Borgerson 2009), but rather the opposite: if we want impartial clinical trials, so that nobody can exploit their inherent uncertainty for their own purposes, we need as many debiasing procedures as possible. Randomization is just one of them and independently of our probabilistic convictions we have good reasons to use it in contexts where biases may interfere (Berry and Kadane 1997)

4. RANDOMIZATION IN FIELD TRIALS

The assessment of public policy programs through large-scale randomized trials is already several decades old (the 1968 New Jersey negative income tax experiment is considered a pioneering example). Usually the interventions assessed deal with one or another aspect of the welfare of large populations and testing them is expensive, though the cost of the actual implementation of the program would be significantly more so. Around 200 randomized field trials were run in the United States between 1960 and 1995 (Orr 1999), with more or less convincing results.

In the last decade, there has been an explosion of interest in randomized social experiments among development economists. Several programs for improving health or education, different microfinance and governance schemes have been tested in a number of developing countries. A success story is the PROGRESA program implemented in Mexico in 1998. PROGRESA aimed at improving school performance through a system of direct transfers conditional on both family income, school attendance and preventive health measurements. The amount of the allocation, received directly by the mothers, was calculated to match the salary of a teenager. In order to test the effects of PROGRESA (and with a view to secure its continuation if there was a change in government), a team at the Ministry chose 506 villages implementing PROGRESA in half of them, selected at random. The data showed an increase in teenager enrollment in secondary education significantly higher in the experimental group, with concomitant improvements in the community health. The experiment was considered convincing enough to ground the extension of the scheme to more than thirty countries.

The boom of field experiments in development economics may owe something to their costs: in developing countries, the costs for running these programs are significantly lower than, say, in the United States and non-governmental organizations can implement it in a quick and efficient manner. But there is also a sense of political opportunity among these social experimentalists. A leading one, Esther Duflo, puts it as follows: just as RCTs brought about a revolution in medicine, randomized field trials can do the

same for the assessment of our education and health policies in fighting poverty (Duflo 2010, p. 17).

Nonetheless, Duflo acknowledges the many methodological pitfalls randomized field experiments can incur. We will only examine here a single controversial aspect of randomization in experiments of these sort, analyzing the arguments of James Heckman, on the one hand, and Duflo and some of her coauthors, on the other, drawing on the discussion of the previous three sections. That is, to what extent can randomization secure causal conclusions and an impartial test?

In 1992, Heckman published a seminal paper containing "most of the standard objections" against randomized experiments in the social sciences. Heckman focused on the non-comparative evaluation of social policy programs, where randomization simply decided who would join them (without allocating the rest to a control group). He presented a semi-formal analysis of the possible effects of randomization on the behavior of potential participants –i.e., the Hawthorne effects it may prompt[20]. In this view, randomization might prevent the experimenters from incurring in selection biases, but it may elicit a different bias on the experimental subjects, who might have behaved differently, if joining the program had not required "a lottery". Randomization would thus interfere with the decision patterns (the causes of action) presupposed in the program under evaluation. Let us present Heckman's case with the responses.

Let D represents participation in a program. Y represents the outcome of this participation. These two variables are related as follows

$$Y = Y_1 \text{ if } D = 1 \quad \text{[The outcome of participating]}$$

$$Y = Y_0 \text{ if } D = 0 \quad \text{[The outcome of not participating]}$$

We presume that the value of Y_0 and Y_1 is causally determined by some umbrella variables X_0 and X_1:

$$Y_1 = g_1(X_1)$$

$$Y_0 = g_0(X_0)$$

[20] Hawthorne effect was first noticed in a study of the productivity of workers under different environmental conditions in the beginning of the 20th century. In a series of experiments that took place at the Hawthorne Works, researchers noted how workers reacted positively to treatments designed to constitute obstacles to productivity, such as poor lighting. The findings were later on attributed to the fact that workers reacted positively to the attention that they received from researchers, and the notion of Hawthorne effect was coined.

If we are evaluating a training program, and Y_1 is the outcome attained by the participants, we may presume it to be determined by their previous education, age, etc. (X_1). Participation in the program is determined in turn by another umbrella variable Z, with a set of values Ψ:

If ; otherwise $D = 0$

For instance, participation may depend on certain values of income, employment, etc., all captured by Z. The collection of explanatory variables in the program assessment is thus $C = (X_0, X_1, Z)$: the outcome depends on certain antecedent factors (captured by X_i) and on participation (Z). We usually do not have full information about C: the available information is represented by C_a. If we conduct an experiment to assess this program, we try to determine the joint probability distribution of Y_1, Y_0, D conditional of a particular value of $C_a = c_a$.

$$F(y_0, y_1, d \mid c_a)$$

In order to make his first objection, Heckman suggests we should distinguish between regular participation in a program (captured by D) and participation in the program in an experimental regime, where participation is randomized. This is captured by a second variable D^*.

$D^* = 1$ if the participant applied and accepted in a regime of random selection

$D^* = 0$ otherwise.

If p is the probability of being accepted in the program after randomization, the possibility of testing the program through randomized tests depends on the following assumption:

$$\Pr(D = 1|c) = \Pr(D^* = 1|c, p)$$

In other words, we need to assume that:

(a) Randomization does not influence participation

(b) If it does, the effect is the same for all the potential participants.

(c) Or, if different, it does not influence their decision to take part in the program.

Heckman's main objection is that randomization tends to eliminate risk-averse persons. This is only acceptable if risk aversion is an irrelevant trait for the outcome under investigation – i.e. it does not feature in C. However, even if irrelevant, it compels experimenters to deal with bigger pools of potential participants in order to meet the desired

sample size, so the exclusion of risk-averse subjects does not disrupt recruitment. But bigger pools may affect in turn the quality of the experiment, if it implies higher costs. One way or another, argues Heckman, randomization is not neutral regarding the results of the experiment. Let us now present the response of Duflo and her coauthors.

According to Banerjee and Duflo, we can avoid Hawthorne effects if we either disguise or hide randomization. Both solutions are feasible in many programs if we are conducting the experiment in a developing country. As to the former, randomization can be disguised as a lottery by which the scarce resources of the program are allocated. If the potential participants perceive this lottery as fair, it may not dissuade them from taking part in it. The fairness of lotteries as allocating procedures can be certainly defended on theoretical grounds (Stone 2007) and we know there is empirical evidence about the acceptability of unequal outcomes when they come from a lottery perceived as fair (Bolton *et al.* 2005). However, not everybody likes lotteries, even fair ones: e.g., there are surveys showing that people oppose the use of lotteries by colleges and universities in order to choose which students are admitted (Carnevale & Rose 2003)

It is an empirical question to be solved on a case by case bases if disguising randomization as a lottery influences participation. Banerjee and Duflo certainly acknowledge that even fair lotteries can provoke a Hawthorne effect depending on the way they are presented: if the participants in the control group are told that the experimental treatment will be available to them in the future (once the resources are gathered), this may affect their willingness to participate or their compliance.

Hiding randomization from participants seems a more effective strategy. As Banerjee and Duflo observe, "ethics committees typically grant an exemption from full disclosure until the end-line survey is completed, at least when the fact of being studied in the control group does not present any risk to the subject" (2009, p. 20). Participants in the experimental group will not know how they got involved and those in the control group may never know they have been excluded. If these latter are in different villages, as it often happens in trials run in developing countries, they may not get to know about the experimental treatment. But, again, we need to verify they actually ignore it: as drug trials in developing countries illustrate, once access to experimental treatments becomes a politically contentious issue within the country, we cannot take ignorance for granted (Macklin 2004, Petryna 2009).

A minimum of cooperation is necessary in a trial. The participants in medical experiments usually do not understand randomization well and they do not like it much (Featherstone & Donovan 2002): their compliance is usually explained by their lack of alternatives to get access to experimental treatments. But this is often depends on the social organization of the patients. The testing of early anti-AIDS treatments in the USA, documented by Epstein 1996, illustrates this point: the participants wanted to have experimental treatments and not placebos so they resorted to all sort of strategies to make sure they got the former, drawing on their connections in the gay activism networks. Many abstained from taking part in trials if they didn't think the drug was promising enough (in order to remain "clean" and eligible for further tests); those who participated exchanged the pills between them (at the cost of halving the dose) or took them to independent laboratories to verify the active principle. They simply collapsed the trial protocol.

In social trials Hawthorne effects can still take place beyond the recruitment stage: people may behave differently if they realize in which group they are. In order to control for this post-randomization effects, Duflo, Glennerster and Kremer (2007) suggest two additional strategies. The first one is collecting longer run data, once the experiment is terminated, in order to verify whether the interaction with the experimenter was making any difference in the behavior of the participants –e.g., Duflo and Hanna 2006.

Depending on the experiment, we can also try to control for the mechanisms prompting the Hawthorne effects. We can create additional control groups in which we expose the participants to, e.g., the same type of interaction with the experimenter (in the presentation and administration of the treatment) as the experimental group. In a test of a savings program we have three groups: the experimental treatment (the new savings scheme), the control treatment (the regular savings scheme "administered" in ordinary conditions) and a third control group, in which the participants receive a marketing pitch of a standard savings scheme, similar to the one used with the experimental treatment, in order to see whether the pitch per se had any effect. This strategy is most effective when there is an explicit interaction between the experimenter and the participants that we can reproduce independently of the actual treatment.

However, these are just two possible controls on the behavior of the experimental subject, but the possibility of strategic interaction between experimenters and participants can never be ruled out. So far we have considered the question from the standpoint of

causality, but it also affects the impartiality of trials in development economics. As we saw in the previous two sections, when the conclusion of the experiment involves wins and losses to the parties concerned, it is crucial that it appears as a fair test for its results to be publicly accepted. In clinical trials randomization provides a warrant of fairness, but we saw that it is not enough to cope with the many biases that can emerge at the different stages of design and implementation of a trial. In social trials, randomization provides a cue for the participants to trump the impartiality of the test with their own decisions, either for personal ("I cannot be bothered") or political reasons ("I do not want this sort of policy implanted"). Randomization is thus not such a good warrant of impartiality in development experiments. And, given the institutional organization of these trials, we may wonder if there is any.

Randomized evaluations of development policies are still rare and the experimenters involved do not seem to be aware of all the potential sources of bias presented in table 2. But they are clearly aware of the problems of credibility posed by their program assessments, which are partly due to randomization. Organizing a lottery to distribute aid seems to be politically controversial for governments that are expected to serve an entire population. Duflo, Glennerster and Kremer (2007, p. 21). Government-sponsored programs are rare because it is difficult to attain the high level of political consensus required for a successful implementation. Without this consensus, randomized evaluations can be easily prey to the sort of manipulations described above. Non-governmental organizations are more active, because they are interested in finding the most efficient way of spending their (usually scarce) resources and they are comparatively free to choose where and how they distribute them. However, NGOs create their own Hawthorne effects: the culture of the organization implementing the assessment (e.g., the motivation of its employees) may impact on the participants' reaction in a way difficult to replicate in further extensions of the program.

NGOs (or for-profit organizations for that matter) have also a problem of credibility, not unlike the pharmaceutical industry: they usually have a stake in the programs they evaluate (Pritchett 2002). And randomization does not seem to be a good enough warrant of impartiality to convince governments that they can trust an assessment and implement it at a bigger scale. This is probably why Duflo and Kremer (2005, pp. 115-117) advocate the creation of a sort of international "regulatory agency" for development policies. International organizations involved in development should establish an office with the

following mission. It should assess the "ability of the evaluation to deliver reliable causal estimates of the project's impact" and "conduct credible evaluations in key areas (p.115).

In other words, international organizations should provide the impartial expertise required to make the trials credible to the involved parties. Including an assessment of the potential Hawthorne effects prompted by randomization. This is probably the best solution and we think it should be implemented. However, it remains an open question why would the participants in the trial see the international organization as a neutral third party they can trust. Only if they do, we can be certain that the trials it sponsors are a credible source of knowledge about their target population.

REFERENCES

Abraham, John. "Pharmaceuticalization of Society in Context: Theoretical, Empirical and Health Dimensions." *Sociology* 44, no. 4 (2010): 603-622.

Abraham, John, and Courtney Davis. "Drug Evaluation and the Permissive Principle." *Social Studies of Science* 39, no. 4 (2009): 569-598.

Ancker, J.S. , and A. Flanagin. "A Comparison of Conflict of Interest Policies at Peer-Reviewed Journals in Different Scientific Disciplines." *Sci Eng Ethics* 13, no. 2 (2007): 147-57.

Angell, Marcia. *The Truth About the Drug Companies : How They Deceive Us and What to Do About It*. 1st ed. New York: Random House, 2004.

Banerjee, Abhijit V., and Esther Duflo. "The Experimental Approach to Development Economics." *Annual Review of Economics* 1, no. 1 (2009): 151-178.

Bekelman, Justin E., Yan Li, and Cary P. Gross. "Scope and Impact of Financial Conflicts of Interest in Biomedical Research." *JAMA: The Journal of the American Medical Association* 289, no. 4 (2003): 454-465.

Berry, Scott M., and Joseph B. Kadane. "Optimal Bayesian Randomization." *Journal of the Royal Statistical Society. Series B (Methodological)* 59, no. 4 (1997): 813-819.

Bero, Lisa A., and Drummond Rennie. "Influences on the Quality of Published Drug Studies." *International Journal of Technology Assessment in Health Care* 12, no. 02 (1996): 209-237.

Borgerson, Kirstin. "Valuing Evidence Bias and the Evidence Hierarchy of Evidence-Based Medicine." *Perspectives in Biology and Medicine* 52, no. 2 (2009): 218-233.

Biddle, Justin. "Lessons from the Vioxx Debacle: What the Privatization of Science Can Teach Us About Social Epistemology." *Social Epistemology: A Journal of Knowledge, Culture and Policy* 21, no. 1 (2007): 21 - 39.

Boldrin, Michele, and David K. Levine. *Against Intellectual Monopoly*. New York: Cambridge University Press, 2008.

Bolton, Gary E, Jordi Brandts, and Axel Ockenfels. "Fair Procedures: Evidence from Games Involving Lotteries." *The Economic Journal* 115, no. 506 (2005): 1054-1076.

Carnevale, Anthony Patrick, Stephen J. Rose, *Socioeconomic Status, Race/Ethnicity, and Selective Admissions*. The Century Foundation, 2003.

Carpenter, Daniel P. *Reputation and Power : Organizational Image and Pharmaceutical Regulation at the Fda*. Princeton: Princeton University Press, 2010.

Carpenter, Daniel, and Colin Moore. "Robust Action and the Strategic Use of Ambiguity in a Bureaucratic Cohort: Fda Scientists and the Investigational New Drug Regulations of 1963." In *Formative Acts*, ed. Stephen Skowronek and Matthew Glassman, 340-362. Philadelphia: University of Pennsylvania press, 2007.

Cartwright, Nancy. "Are Rcts the Gold Standard?" *Biosocieties* 2, no. 1 (2007): 11-20.

_____. "What Are Randomised Controlled Trials Good For?" *Philosophical Studies* 147, no. 1 (2010).

Cartwright, Nancy, and Eileen Munro. "The Limitations of Randomized Controlled Trials in Predicting Effectiveness." *Journal of Evaluation in Clinical Practice* 16, no. 2 (2010): 260-266.

Ceccoli, Stephen J. "The Politics of New Drug Approvals in the United States and Great Britain." Thesis (Ph. D.), Washington University, 1998.

Chalmers, Iain. "Statistical Theory Was Not the Reason That Randomization Was Used in the British Medical Research Council's Clinical Trial of Streptomycin for Pulmonary Tuberculosis." In *Body Counts: Medical Quantification in Historical and Sociological Perspective*, ed. Gérard Jorland, George Weisz and Annick Opinel, 309-334. Montreal: McGill-Queen's Press, 2005.

Doucet, M, and S Sismondo. "Evaluating Solutions to Sponsorship Bias." *Journal of Medical Ethics* 34, no. 8 (2008): 627-630.

Dowling, H. F. "Twixt the Cup and the Lip." *Journal of the American Medical Association* 165, no. 6 (1957): 657-61.

Duflo, Esther. *La Politique De L'autonomie*. Paris: Seuil, 2010.

_____. *Le Développement Humain*. Paris: Seuil, 2010.

Duflo, Esther, Rachel Glennerster, and Michael Kremer. "Using Randomization in Development Economics Research: A Toolkit." C.E.P.R. Discussion Papers, CEPR Discussion Papers: 6059, 2007.

Duflo, Esther, and Rema Hanna. "Monitoring Works: Getting Teachers to Come to School." C.E.P.R. Discussion Papers, CEPR Discussion Papers: 5426, 2006.

Duflo, Esther, and Michael Kremer. "Use of Randomization in the Evaluation of Development Effectiveness." In *Evaluating Development Effectiveness*, ed. George Keith Pitman, Osvaldo N. Feinstein and Gregory K. Ingram, 205-232: World Bank Series on Evaluation and Development, vol. 7. New Brunswick, N.J. and London: Transaction, 2005.

Epstein, Steven. *Impure Science. Aids and the Politics of Knowledge*. Berkeley-Los Angeles: University of California Press, 1996.

Featherstone, Katie, and Jenny L. Donovan. ""Why Don't They Just Tell Me Straight, Why Allocate It?" the Struggle to Make Sense of Participating in a Randomised Controlled Trial." *Social Science & Medicine* 55, no. 5 (2002): 709-19.

Fisher, Jill A. *Medical Research for Hire : The Political Economy of Pharmaceutical Clinical Trials*. New Brunswick, N.J.: Rutgers University Press, 2009.

Hackshaw, Allan K. *A Concise Guide to Clinical Trials*. Chichester, UK ; Hoboken, NJ: Wiley-Blackwell/BMJ Books, 2009.

Heckman, James J. "Randomization and Social Policy Evaluation." In *Evaluating Welfare and Training Programs*, ed. Charles F. Manski and Irwin Garfinkel, 201-230: Cambridge and London: Harvard University Press, 1992.

Jadad, Alejandro R., R. Andrew Moore, Dawn Carroll, Crispin Jenkinson, D. John M. Reynolds, David J. Gavaghan, and Henry J. McQuay. "Assessing the Quality of

Reports of Randomized Clinical Trials: Is Blinding Necessary?" *Controlled Clinical Trials* 17, no. 1 (1996): 1-12.

Jain, Shaili. *Understanding Physician-Pharmaceutical Industry Interactions*. Cambridge [England] ; New York: Cambridge University Press, 2007.

Krimsky, Sheldon. *Science in the Private Interest : Has the Lure of Profits Corrupted Biomedical Research?* Lanham: Rowman & Littlefield Publishers, 2003.

Lasagna, L. "Gripesmanship: A Positive Approach." *Journal of Chronic Diseases* 10 (1959): 459-68.

Levi, Isaac. "Direct Inference and Randomization." *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association* 2 (1982): 447-463.

Lexchin, Joel, Lisa A Bero, Benjamin Djulbegovic, and Otavio Clark. "Pharmaceutical Industry Sponsorship and Research Outcome and Quality: Systematic Review." *BMJ* 326, no. 7400 (2003): 1167-1170.

Macklin, Ruth. *Double Standards in Medical Research in Developing Countries* Cambridge Law, Medicine, and Ethics ; 2. Cambridge, UK ; New York, NY: Cambridge University Press, 2004.

Marks, H. M. *The Progress of Experiment. Science and Therapeutic Reform in the United States, 1900-1990*. N. York: Cambridge University Press, 1997.

Marks, Harry. "Trust and Mistrust in the Marketplace: Statistics and Clinical Research, 1945-1960." *History of Science* 38 (2000): 343-355.

Meldrum, Marcia Lynn, *Departures from the Design: The Randomized Clinical Trial in Historical Context, 1946-1970*, Ph.D. Dissertation, State University of New York, 1994.

Montori, Victor M, Roman Jaeschke, Holger J Schanemann, Mohit Bhandari, Jan L Brozek, P J Devereaux, and Gordon H Guyatt. "Users' Guide to Detecting Misleading Claims in Clinical Research Reports." *BMJ* 329, no. 7474 (2004): 1093-1096.

Murphy, Timothy F. *Case Studies in Biomedical Research Ethics* Basic Bioethics. Cambridge, Mass.: MIT Press, 2004.

Orr, Larry L. *Social Experiments : Evaluating Public Programs with Experimental Methods*. Thousand Oaks, Calif.: Sage Publications, 1999.

Petryna, Adriana. *When Experiments Travel : Clinical Trials and the Global Search for Human Subjects*. Princeton: Princeton University Press, 2009.

Piantadosi, Steven. *Clinical Trials : A Methodological Perspective*. 2nd ed. Wiley Series in Probability and Statistics. Hoboken, N.J.: Wiley-Interscience, 2005.

Pocock, Stuart J. *Clinical Trials : A Practical Approach* A Wiley Medical Publication. Chichester [West Sussex] ; New York: Wiley, 1983.

Pritchett, Lant. "It Pays to Be Ignorant: A Simple Political Economy of Rigorous Program Evaluation." *Journal of Policy Reform* 5, no. 4 (2002): 251-269.

Sacristán, J.A., E. Bolanos, J.M. Hernández, J. Soto, and I. Galende. "Publication Bias in Health Economic Studies." *Pharmacoeconomics* 11 (1997): 289-292.

Sheps, M. C. "The Clinical Value of Drugs: Sources of Evidence." *American Journal of Public Health & the Nation's Health* 51, no. 5 (1961): 647-54.

Sismondo, Sergio. "How Pharmaceutical Industry Funding Affects Trial Outcomes: Causal Structures and Responses." *Social Science & Medicine* 66, no. 9 (2008): 1909-1914.

Stone, Peter. "Why Lotteries Are Just?" *The Journal of Political Philosophy* 15, no. 3 (2007): 276-295.

Tavris, Carol, and Elliot Aronson. *Mistakes Were Made (but Not by Me) : Why We Justify Foolish Beliefs, Bad Decisions, and Hurtful Acts*. 1st ed. ed. Orlando, Fla.: Harcourt, 2007.

Thompson, D. "Understanding Financial Conflicts of Interest." *New England Journal of Medicine* 329, no. 8 (1993): 573-576.

Urbach, Peter. "Randomization and the Design of Experiments." *Philosophy of Science. JE* 85, no. 52 (1985): 256-273.

_____. "A Reply to Mayo's Criticisms of Urbach's "Randomization and the Design of Experiments"." *Philosophy of Science. Mr* 91 (1991): 125-128.

_____. "Reply to David Papineau." *British Journal for the Philosophy of Science* 45, no. 2 (1994): 712-715.

Worrall, John. "What Evidence in Evidence-Based Medicine?" *Philosophy of Science* 69, no. 3 Supplement (2002): S316-S330.

_____. "Evidence in Medicine and Evidence-Based Medicine." *Philosophy Compass. N* 2, no. 6 (2007): 981-1022.

_____. "Why There's No Cause to Randomize." *British Journal for the Philosophy of Science. S* 58, no. 3 (2007): 451-488.

_____. "Evidence and Ethics and Medicine." *Perspectives in Biology and Medicine* 51, no. 3 (2008): 418-431.