Lafollette, Hugh, and Niall Shanks. 1996. *Brute science: Dilemmas of animal experimentation*. New York: Routledge.
Levine, Phillip, Douglas Staiger, Thomas Kane, and David Zimmerman. 1999. *Roe v. Wade* and American fertility. *American Journal of Public Health* 89(2): 199–203.
Levitt, Steven. 2004. Understanding why crime fell in the 1990s: Four factors that explain the decline and six that do not. *Journal of Economic Perspectives* 18(1): 163–190.
Levitt, Steven. 2005. Abortion and crime: Who should you believe? *Freakonomics Blog*. http://www.freakonomics.com/2005/05/15/abortion-and-crime-who-should-you-believe/
Little, Daniel. 1992. *Varieties of social explanation*. Boulder: Westview Press.
Little, Daniel. 1998. *Microfoundations, method, and causation*. New Brunswick: Transaction Publishers.
Matejcek, Zdenek, Zdenek Dytrych, and Vratislav Schuller. 1988. In *The Prague cohort through age nine*, ed. H.P. David et al., 53–86. New York: Springer.
Pearl, Judea. 2000. *Causality: Models, reasoning, and inference*. Cambridge: Cambridge University Press.
Pearl, Judea, and Daniel Bareinboim. 2011. Transportability across studies: A formal approach. UCLA Cognitive Systems Laboratory, Technical Report (R-372), San Francisco.
Rasanen, Pirkko, Helina Hakko, Matti Isohanni, Sheilagh Hodgkins, Marjo-Riita Jarvelin, and Jari Tiihonen. 1999. Maternal smoking during pregnancy and risk of criminal behavior among adult male offspring in the northern Finland 1966 birth cohort. *The American Journal of Psychiatry* 156(6): 857–862.
Reiss, Julian. 2007. Do we need mechanisms in the social sciences? *Philosophy of the Social Sciences* 37(2): 163–184.
Reiss, Julian. 2010. Review: Across the boundaries: Extrapolation in biology and social science. *Economics and Philosophy* 26: 382–390.
Reyes, Jessica. 2007. Environmental policy as social policy? The impact of childhood lead exposure on crime. *The B.E. Journal of Economic Analysis & Policy* 7(1): Article 51.
Steel, Daniel. 2004. Social mechanisms and causal inference. *Philosophy of the Social Sciences* 34 (1): 55–78.
Steel, Daniel. 2008. *Across the boundaries: Extrapolation in biology and social science*. New York: Oxford University Press.
Steel, Daniel. 2011. Causality, causal models, and mechanisms. In *The Sage handbook on the philosophy of social science*, ed. Ian C. Jarvie and Jesús Zamora-Bonilla, 288–304. London: Sage.
Truman, Jennifer, and Michael Rand. 2010. Criminal victimization, 2009. *Bureau of Justice Statistics Bulletin* (NCJ 231327). Washington, DC: United States Department of Justice, Bureau of Justice Statistics.
Wadsworth, Tim. 2010. Is immigration responsible for the crime drop? An assessment of the influence of immigration on changes in violent crime between 1990 and 2000. *Social Science Quarterly* 91(2): 531–553.

Chapter 11
Causality, Impartiality and Evidence-Based Policy

David Teira and Julian Reiss

Abstract The overall aims of this chapter are to compare the use of randomised evaluations in medicine and economics and to assess their ability to provide impartial evidence about causal claims. We will argue that there are no good reasons to regard randomisation as a sine qua non for good evidential practice in either science. However, in medicine, but not in development economics, randomisation can provide impartiality from the point of view of regulatory agencies. The intuition is that if the available evidence leaves room for uncertainty about the effects of an intervention (such as a new drug), a regulator should make sure that such uncertainty cannot be exploited by some party's private interest. We will argue that randomisation plays an important role in this context. By contrast, in the field evaluations that have recently become popular in development economics, subjects have incentives to act strategically against the research protocol which undermines their use as neutral arbiter between conflicting parties.

List of Abbreviations

FDA	Food and Drug Administration
NGO	Non-Governmental Organisation
RCT	Randomised Clinical Trial
RFT	Randomised Field Trials

D. Teira (✉)
Departamento de Lógica, Historia y Filosofía de la ciencia, UNED, Paseo de Senda del rey 7, 28040 Madrid, Spain
e-mail: dteira@fsof.uned.es

J. Reiss
Department of Philosophy, Durham University, 50 Old Elvet, Durham DH1 3HN, UK
e-mail: julian.reiss@durham.ac.uk

H.-K. Chao et al. (eds.), *Mechanism and Causality in Biology and Economics*, History, Philosophy and Theory of the Life Sciences 3, DOI 10.1007/978-94-007-2454-9_11,

1 Introduction

Randomisation, the assignment of experimental subjects to treatment groups by means of a random number generator, was first systematically applied in psychic research in the late nineteenth century and became popular in statistics after Ronald Fisher advocated its use in 1926 (Hacking 1988). In medicine and development economics, the two sciences we will focus on in this chapter, randomised trials are now widely regarded as the 'gold standard' of evidence. The overall aims of this chapter are to compare the use of randomised evaluations in these two sciences and to assess their ability to provide impartial evidence about causal claims. In short, we will argue that there are no good reasons to regard randomisation as a sine qua non for good evidential practice in either science. However, in medicine, but not in development economics, randomisation can provide impartiality from the point of view of regulatory agencies. The intuition is that if the available evidence leaves room for uncertainty about the effects of an intervention (such as a new drug), a regulator should make sure that such uncertainty cannot be exploited by some party's private interest. We will argue that randomisation plays an important role in this context. By contrast, in the field evaluations that have recently become popular in development economics, subjects have incentives to act strategically against the research protocol which undermines their use as neutral arbiter between conflicting parties.

2 Background: Randomised Clinical Trials as a Public Policy Tool

Randomised clinical trials (RCTs) are medical experiments in which alternative treatments for a condition are administered to at least two groups of patients in order to see which one is the safest and most effective for future cases. Unlike other experimental designs in medicine and elsewhere, RCTs have achieved some public notoriety throughout the last five decades thanks to the role they play in pharmaceutical regulation. The commercial distribution of novel drugs will only be authorised by regulatory agencies such as the American Food and Drug Administration (FDA) if their safety and efficacy is proved in two RCTs.

RCTs come to solve a problem in public policy: by their own means, consumers cannot ascertain the quality of a drug, either by simple inspection of their appearance (shape, size, smell, etc.) or by their price. Depending on the circumstances of the patient, the natural rate of variability of their effects (positive or negative) prevents a reliable assessment on the basis of individual experience alone. When buyers or sellers cannot directly determine the quality of a good or service, economic *adverse selection* can lead to the elimination of all trade in a market

(Akerlof 1970; Wilson 2008): putting it very simply, if consumers know that there is a percentage of cheap, bad-quality drugs, they will be reluctant to pay the price requested for good-quality compounds, and the producers of the latter may end up leaving the market. Regulation may be justified to remedy this type of market failure whose consequences can be fatal for the consumers (because they need the good-quality drugs).

RCTs are one way to provide the grounds for an evidence-based pharmaceutical policy: the regulator will make a decision on the marketing of a new drug depending on the evidence RCTs yield about its safety and efficacy. One may wonder, however, why RCTs are regarded as credible, given the conflicts of interest that pervade the pharmaceutical markets. Historically, physicians, pharmacists and patients have supported their favourite treatments, seeking whatever evidence confirmed their views and questioning, with the same passion, the quality of any piece of adverse evidence. Why should they now accept RCTs?

A standard sociological response is because in democratic societies RCTs provide an appearance of *mechanical objectivity* that seems more acceptable than mere expert clinical judgement: the statistical apparatus underlying RCTs proceeds impartially, impervious to the particular interests that may bias the judgement of the individual expert. But, so the standard response continues, mechanical objectivity is a mere appearance caused by numbers whose statistical justification lay audiences cannot grasp (see, for instance, Porter 1995; Marks 1997).

In these sociological accounts 'mechanical objectivity' contrasts with 'expert judgement'. The so-called evidence-based medicine (e.g. Sackett et al. 1996) is a paradigmatic example for the perennial attempts to replace the latter by the former. In pre-evidence-based medicine, the standard approach to assessing the efficacy of new treatments was heavily influenced by clinicians' judgements. But clinicians, like all experts, may be inattentive, ill informed, partial (to this or that therapy) or otherwise biased. Moreover, an expert's decision is not transparent to outsiders (in this case, patients). There are therefore good reasons to limit the influence of the clinician's judgement to a minimum and replace it with 'objective evidence'. Objective evidence is sometimes called 'mechanical' when it is produced by mechanical methods such as RCTs. An RCT is a mechanical method in that its implementation follows strict and explicit rules – divide the test population into two groups by means of a random allocation mechanism, blind subjects and treatment administrators, follow specific stopping rules, etc. Unlike expert judgements, such mechanical rules are transparent. This means that they can be publicly debated, scrutinised and criticised.

Over the last 10 years, philosophers of science such as Nancy Cartwright and John Worrall have challenged the epistemic foundations for RCTs (e.g. Cartwright 2007; Cartwright and Munro 2010; Worrall 2002, 2007). They appraise RCTs as tools for causal inference. In their – philosophers' – approach, impartiality is at best a by-product of causal analysis: if one can establish objectively that a drug is effective in curing a given condition, this judgement is independent from whatever

interest there might be at stake in the experiment. Both Cartwright and Worrall conclude that RCTs are not completely reliable tools for causal inference, and therefore, we should consider alternative sources of evidence for our regulatory decisions, including expert judgement. In the following section, we will examine Cartwright's criticism in more detail.

3 RCTs as 'Gold Standard' of Evidence for Causal Claims

According to Nancy Cartwright, RCTs are just one method among others for warranting causal claims. In her terminology, it is a 'clinching' method in that it proceeds in a deductive fashion: *if* its assumptions are met and the observable evidence is positive, *then* we can safely affirm the causal claim. However, the premises are restrictive, and therefore, the range of conclusions narrow (Cartwright 2007).

Ideal RCTs test causal claims about the narrow efficacy claims of drugs in a given population following Mill's method of difference. Given an observed 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, all causally relevant factors other than T are equally distributed. (This is what randomisation is supposed to achieve; more on that later.) Therefore, the observed difference in O must be an effect of T. To show that the effectiveness claim follows deductively, we need a number of further assumptions.

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 Patrick Suppes' probabilistic theory of causality which states that for an event-type T preceding event-type O in a population Θ,

$$\text{T causes O in } \Theta \text{ iff } P(O/T\&K_i) > P(O/\neg T\&K_i)$$
$$\text{for some subpopulation } K_i, \text{ with } P(K_i) > 0.$$

Cartwright further assumes that the individuals in the sample are all governed by the same causal structure CS, 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'. 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 T and O. 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 ideal RCTs can clinch. However, according to Cartwright, we need further assumptions still if we want to generalise this conclusion to some target population Θ.

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:

(a) 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.
(b) The causal structure and the probability measure are the same in that subpopulation of Θ as it is in that subpopulation φ.

The warrant for these assumptions too is supposed to come from randomisation, but we cannot judge whether a group of patients constitutes a random sample without a previous idea of what factors are to be equally represented (Cartwright 2007, p. 18). In a trial, we want to form, on the one hand, two treatment groups that are balanced with respect to known relevant prognostic factors. On the other hand, we want to avoid unknown confounders to affect the result. Randomisation supposedly helps us in achieving both goals, but it is neither necessary nor sufficient to that effect. By sheer chance, a random allocation may yield an unbalanced distribution of the prognostic factors between the treatment groups (these are called 'baseline imbalances'). This may bias the comparison between treatments and invalidate the experimental results, and when imbalances occur, trialists usually try to correct them (e.g. by repeating the randomisation). Unknown confounders may differentially influence the outcome in one of the groups even after a randomised allocation of treatments. Further randomisations at each step in the administration of the treatment (e.g. which nurse should administer the treatment today?) may avoid such interferences, but this is impracticable. Declaring such disturbances as negligible, as many experimenters do, lacks any justification in the assumed statistical methodology (Urbach 1985; Worrall 2007).

Both the correction of imbalanced allocations and the decision to randomise at different stages of the trial beyond the allocation of treatments require extra-statistical expert judgement. Against the ideal of mechanical objectivity, we need an expert who can handle different sources of evidence other than the trial to justify the acceptance of assumptions (a) and (b). More precisely, we need someone who can certify that randomisation, the main warrant of (a) and (b), has indeed worked. Without this judgement, subjective and intransparent as it may be, we cannot safely generalise the conclusions of the trial to its target population, i.e. ascertain its external validity. Expert judgements are naturally fallible too, but, according to Cartwright (2007, p. 19), to rely on mechanical methods without expertise and watch out for failures is no satisfactory response.

4 External Validity and Impartiality in Regulatory RCTs

Today, FDA is probably the institution that makes the most systematic use of RCTs for regulatory purposes in the world, but has not always been so. Between 1900 and 1950, expert clinical judgement was the main criterion in the assessment of the properties of pharmaceutical compounds in the United States as well as in other countries such as Britain. An experienced clinician would administer the drug to a series of patients he would consider likely to benefit. His or her conclusions would be presented as a case report, informing of the details of each patient's reaction to the treatment. The regulatory authorities in the United States and Britain arranged official drug testing depending on the standards adopted by the research community within their respective medical professions. Until the 1960s, regulatory decisions were fundamentally based on expert judgements of this sort. Clinical judgement came to be discredited in the United States because a group of methodologically minded pharmacologists took over the FDA and imposed their views on the superiority of RCTs through regulatory means. This was the triumph of mechanical objectivity against expert judgement.

During the 1960s and 1970s, RCTs became mandatory for regulatory decisions in different degrees. In the United States, before the 1960s, the FDA was entitled only to test the safety but not the efficacy of pharmaceutical compounds. In the late 1950s, there were voices in the FDA demanding stricter testing standards linking safety and efficacy, under increasing public mistrust in the pharmaceutical industry, prompted in part by the thalidomide scandal.

Under the trade name *Contergan*, one million West Germans consumed thalidomide as a sedative in the early 1960s and many more people around the world after that. Reports showing an association between the drug and peripheral neuropathy were soon published in medical journals. Later reports of serious birth defects when the drug was consumed by pregnant women surfaced. Only then did the manufacturer withdraw the drug from European markets. Eight thousand babies had been already born with severe deformities. At that point, there was no clear regulatory standard about the safety of a compound, neither in the United States nor in Europe.

The thalidomide scandal gave them the opportunity to put their views in effect in 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 (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 prior trial or posterior confirmatory trial). It has been claimed, correctly in our view, that this set of regulations created the modern clinical trial industry (Carpenter and Moore 2007). In the following three decades, pharmaceutical funding would boost the conducting of RCTs (by the thousands) in the United States and abroad.

The regulatory approach of the FDA surely constitutes a canonical instance of an evidence-based policy or, more precisely, an RCT-based policy. It is worth noting

that the FDA does not take the external validity of an RCT for granted. A drug trial is usually divided in four phases. Phase I focuses on finding the appropriate dosage in a small group of healthy subjects (20–80 patients); toxicity and other pharmacological properties of the drug are examined. In phase II, between 100 and 200 hundred patients are closely monitored to verify the treatment effects. If treatment effects are detected, a third phase involving a substantial number of patients begins in which the drug is compared to the standard treatment. This is usually referred to as 'the' RCT. If the new drug proves to be at least as good as the existing therapies and the pharmaceutical authorities approve its commercial use, phase IV starts: the drug is freely prescribed and sold; adverse effects are monitored and morbidity and mortality studies are undertaken.

In other words, the FDA, as other regulatory agencies, does not take the external validity of the RCTs for granted when it approves a new substance. In the post-market surveillance phase IV, the FDA collects adverse event reports from various sources and conducts epidemiological studies to assess their relevance, keeping track of the validity of the results of their trials in the general population. 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 large: apart from monitoring adverse reports, the agency also has to consider issues in labelling, advertising or the inspection of production and storage facilities, to name but a few. Hence, one should appraise the figures collected at this fourth phase *cum mica salis*. But we think they are relevant in the context of our analysis of the reliability of RCTs.

Changes in drug labelling constitute a first approximation to the number of minor or major failures at phase III trials. 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 (advertising and prescription are only lightly regulated in the United States as compared to Europe). As long as the label records 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 (Fig. 11.1):

Carpenter (2010, p. 623) summarises it as follows: from 1980 to 2000, the average new molecular entity received five labelling revisions after approval, about one for every 3 years of marketing after approval. Only one in four drugs had no labelling revisions at all. The data are obviously too coarse to decide what went wrong, if anything, in the phase III RCTs. Several explanations are possible: for instance, the trials might have been too brief to detect adverse effects (e.g. toxicity or cardiovascular events). In the context of Cartwright's analysis, we may suspect that the patients' sample might have been unrepresentative of the patients that finally used the therapy. In a rough sense, the list of prognostic factors and potential confounders used to define the eligibility criteria was incomplete and randomisation could not correct this flaw. If we take external validity in an equally rough sense, Carpenter's data would suggest that Cartwright points out correctly the

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

Fig. 11.1 Drug changes requiring a supplemental new drug application, 1970–2006 (Carpenter 2010, p. 613)

limitations of causal inference in RCTs: the two phase III RCTs that granted the approval of the drug do not usually capture the full range of effects of a drug.

However, it is useful to compare these figures with drug withdrawals. We should always bear in mind that phase III trials are testing the safety and efficacy of a compound, but not their full range of effects, which are only seen in phase IV. The figures should be taken again with caution, since, as Carpenter (2010, ch. 9) warns, the negotiation of each withdrawal depends on a number of circumstances outside and inside the agency, among which a prominent one is the time constraints for the review process (cf. Carpenter 2010). However, very few compounds have been withdrawn from the market in the United States during the last five decades for lack of safety or efficacy after receiving the authorisation of the FDA: if we exclude the drugs approved just before the new legal deadline established in 1992, for which security issues seem to be more prominent, between 1993 and 2004 only 4 out of the 211 authorised drugs were withdrawn.

If we thus take label revisions and market withdrawals as rough indexes of the external validity of the regulatory trials approved by the FDA, we may conclude that the procedure is not foolproof (in the sense of anticipating every safety threat a drug may pose), but that it does not fare completely badly either. Its main effects are reasonably well anticipated. Of course, this is a black box argument: we know that the four-phase regulatory system at the FDA screens off dangerous compounds, but perhaps this is just because the pharmaceutical industry does not dare to submit any potentially dangerous new compound. Assuming that the FDA system works (and very few people question that it does), RCTs certainly do not explain its success

alone. Lots of formal and informal causal knowledge are acquired in the first two phases, and it is put to test not only in the RCTs, but by the subsequent epidemiological surveillance, if the drug is approved.

Expert judgement contributes to all the four phases, and even if the decision to authorise a drug is taken at the third phase, the decision is not *mechanical*, precisely because the external validity of a trial cannot be taken for granted. But this is something pharmacologists have known right from the beginning: the regulatory system at the FDA was established to deal with causal uncertainty. During most of the second half of the twentieth century, pharmaceutical research advanced through a so-called molecular lottery: compounds were synthesised and tested on animals without any clear theoretical guidance, and even when they had an interesting therapeutic effect, there often was no grasp of the precise mechanism responsible for it. RCTs allowed pharmacologists to deal with this causal uncertainty about drugs, focusing on the probability of attaining certain treatment outcomes under a given range of prognostic factors.

But if RCTs rely so crucially on expert judgement, we may wonder why they were considered an improvement over the older methodology. One of us has defended elsewhere that all the involved parties explicit sought a testing methodology with warrants of impartiality against the potential conflicts of interest arising in the trial (Teira 2011a, b). The 1950s saw a boom in industrial drug production (some of which were 'wonder drugs' such as 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). For therapeutic reformers, RCTs with their 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, pp. 460–461). Adopting RCTs as a regulatory tool allowed the FDA to justify the impartiality of their decisions about treatments before patients, physicians and the pharmaceutical industry (Teira 2011a, b).

In this respect, randomisation was considered more as a debiasing procedure than a tool for causal inference. Randomisation prevents researchers from allocating treatments to patients according to their personal interests, so that the healthiest patients get the researcher's favourite therapy. As mentioned above, such unbalanced allocations can happen nonetheless by chance. But randomisation is still a warrant that the allocation was not done on purpose with a view to promoting somebody's interests. A priori, the experimental procedure is impartial with respect to the interests at stake.

Of course, in Cartwright's causal approach, randomisation would just be a tool for controlling the probabilistic dependences arising from selection biases in the experimental and the control group, making sure that they are the same in both treatment groups except for the treatment. To understand randomisation in this way is perfectly appropriate from a methodological point of view, but this is not how it was understood by those who introduced it into the regulatory system. Given how little was known about causation in RCTs at the time, randomisation was not sought

for its contribution to causal analysis but rather for preventing anyone to exploit this uncertainty about causation for her own benefit. The same justification applies to other standard features of RCTs such as masking treatments or having pre-established decision rules for the interpretation of the results (as in significance testing): they provide a priori warrants for the impartiality of the trial.

The research protocol in RCTs constrains expert judgement at various critical points in the generation and interpretation of clinical evidence during a trial. In this sense, pharmaceutical regulation is based as much on the *impartiality* of its evidential base as on the accuracy of its causal conclusions. Perhaps there are other sources of evidence whose external validity is as good as RCTs, but it is an open question if they can be as impartial as the latter. Impartiality is crucial for public policy, and it seems a defensible decision to adopt RCTs instead of mere expert judgement for regulatory purposes: at least, the former provide certain warrants of impartiality.

An obvious objection, of course, is that RCTs are not actually impartial. There is evidence showing, for instance, a *sponsor* bias: industry-funded trials published are more likely to support the experimental therapy than the standard alternative, despite their good methodological quality. We know that RCTs do not control every possible source of bias, e.g. the research protocol does not impose any constraint on the research question that trials should address, and there is no obvious way to decide which one should it be. But rather than an objection against RCTs, it should be a general caveat about every possible source of clinical evidence: the easier it is to manipulate the method, the less we should rely on the evidence it produces for regulatory purposes.

Summing up, despite the problems with their external validity, regulatory RCTs have been reasonably efficient in keeping the American pharmaceutical market clear of unsafe or inefficacious compounds. Moreover, despite all the label revisions, the American public has considered the FDA a reliable regulator (Carpenter 2010), and we contend that this is because RCTs provided a warrant of impartiality for their decision, despite the inherent uncertainty of phase III trials. If we had full information about the effects of a therapy, impartiality would be warranted by default. However, short of that, we need to make sure that a regulatory decision is fair despite their inherent imperfection.

5 The Impartiality of Randomised Field Evaluations

The assessment of public policy programmes through large-scale randomised field evaluations (RFEs) is already several decades old (the 1968 New Jersey negative income tax experiment is often considered to be 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 programme would be significantly more so. Around 200 RFEs 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 RFEs among development economists. Several programmes 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 programme implemented in Mexico in 1998. PROGRESA aimed at improving school performance through a system of direct transfers conditional on 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 a randomly selected half of them. The data showed an increase in teenager enrolment 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 30 countries.

The boom of RFEs in development economics may owe something to their costs: in developing countries, the costs for running these programmes are significantly lower than, say, in the United States, and non-governmental organisations can implement them 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, RFEs can do the same for the assessment of our education and health policies in fighting poverty (Duflo 2010, p. 17).

Nonetheless, Duflo acknowledges that RFEs can involve many methodological pitfalls. Randomisation is a case in point. Field experimentalists in economics expect it to provide a solid foundation for causal analysis, and we have already discussed Cartwright's criticism of this idea. In this section we discuss further whether we can take RFEs in development economics to be impartial. More precisely, our question is whether randomisation is as credible warrant of impartiality in field trials development in economics as it is in medical RCTs. We think not.

Let us present our case by drawing on an analysis due to James Heckman. In 1992, Heckman published a seminal paper containing 'most of the standard objections' against randomised experiments in the social sciences. Heckman focused on the non-comparative evaluation of social policy programmes, where randomisation simply decided who would join them (without allocating the rest to a control group). Heckman claimed that even if randomisation allows the experimenters to reduce selection biases, it may produce a different bias. Specifically, experimental subjects might behave differently if joining the programme did not require 'a lottery'. Randomisation can thus interfere with the decision patterns (the causes of action) presupposed in the programme under evaluation.

Let us briefly present Heckman's semiformal analysis. Let D represent participation in a programme and Y the outcome of participating. These two variables are related as follows:

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

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

Heckman presumes that the values of Y_0 and Y_1 are causally determined by some umbrella variables X_0 and X_1:

$$Y_1 = g_1(X_1)$$

$$Y_0 = g_0(X_0)$$

If we are evaluating a training programme, 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 programme is determined in turn by another umbrella variable Z, with a subset of values Ψ:

$$\text{If } Z \in \Psi, \quad \text{D} = 1; \quad \text{otherwise,} \quad \text{D} = 0$$

For instance, participation may depend on certain values of income, employment, etc., all captured by Z. The collection of explanatory variables in the programme 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 programme, we try to determine the joint probability distribution of Y_1, Y_0, D conditional of a particular value of $C_a = c_a$:

$$F(y_0, y_1, d|c_a).$$

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

$$D^* = 1 \quad \text{if a person is at risk for randomisation.}$$

$$D^* = 0 \quad \text{otherwise.}$$

If p is the probability of being accepted in the programme after randomisation, the possibility of testing the programme through randomised tests depends on the following assumption:

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

In other words, we need to assume either that:

1. Randomisation does not influence participation, or
2. If it does influence participation, the effect is the same for all the potential participants, or
3. If the effect is different, it does not influence their decision to take part in the programme.

Heckman's main objection is that randomisation 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, randomisation is not neutral regarding the results of the experiment.

Heckman's analysis is causal: randomisation can create a self-selection bias distorting the sample of participants on which any inference should rest. We are going to argue that it is impossible to correct this self-selection bias without putting in question the impartiality of the trial. The threat of partiality does not come in this case from the researchers but from the participants themselves. In RFEs, participants may have their own preferences about the compared treatments, and the risk aversion elicited by randomisation is just one of them: people may prefer to make choices about treatments. In order to preserve randomisation and to correct self-selection biases, Duflo and her coauthors try to control the participants' preferences by blinding, i.e. by disguising or hiding the randomised nature of the experiment. We argue that these attempts assume the indifference of the participants regarding the experimental outcome. If the participants have strong preferences about the outcome, masking randomisation will not be enough to enforce the experimental protocol.

According to Banerjee and Duflo (2009), we can avoid the self-selection bias if we either disguise or hide randomisation. Both solutions are feasible in many programmes, at least if we conduct the experiment in a developing country. As to the former, randomisation can be disguised as a lottery by which the scarce resources of the programme 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 that 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: for instance, surveys show that people oppose the use of lotteries by colleges and universities in order to choose which students are admitted (Carnevale et al. 2003).

It is an empirical question to be solved on a case-by-case basis if disguising randomisation as a lottery influences participation. Banerjee and Duflo certainly acknowledge that even fair lotteries can provoke a self-selection 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. In addition, organising a lottery to distribute aid seems to be politically controversial for governments that are expected to serve an entire population (Duflo et al. 2007, p. 21).

Hiding randomisation altogether from participants seems a more effective strategy. As Banerjee and Duflo observe, 'ethics committees typically grant an exemption from full disclosure until the endline 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 the latter live in different villages, as it often happens in trials run in developing countries, they may not get to know about the experimental treatment. In this way a totally different scenario arises: in order to avoid a self-selection bias, we deceive the participants about the comparative structure of the experiment. The experimenters are assuming here that participants only care taking part in a lottery, but, as a matter of fact, they may also have preferences about the treatments tested. They may want to get one rather than the other. Or, if they understand the nature of the experiment, they may even have a favourite treatment that they want to see succeed – e.g. we may well imagine parents preferring direct allocations of cash to send their kids to school rather than paid meals. If these preferences exist, disguising randomisation will only succeed to the extent that the disguise is successful: the participants have been 'blinded' to the comparison, but shall we just assume that such blinding is successful?

There is some evidence that deception in medical trials can fail. Patients have preferences about treatments, and they usually neither understand nor like randomisation (Featherstone and Donovan 2002): their compliance is usually explained by their lack of alternatives to get access to experimental treatments, they would not get the medication outside the trial. And they play by the research protocol only to a point: they try to find out which treatment they are receiving (and if they succeed, this has an effect on the experiment). However, in most medical trials, the researchers have means to make patients comply with the research protocol, e.g. they may mask the treatments well enough for an ordinary participant not to be able to distinguish them. They would need a laboratory. Whether they have access to a laboratory often depends on the social organisation 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 would receive the treatment, 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 thus eligible for other 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 completely undermined the trial protocol.

Drug trials in developing countries illustrate how access to experimental treatments becomes a politically contentious issue within the country (Macklin 2004; Petryna 2009). We can probably expect the same from RFEs in economics: if they address interventions about which the potential participants have preferences, randomisation may elicit a different type of self-selection. Participants may behave differently depending on their taste for a treatment, over-complying if they want it to succeed or the opposite if they want to see it fail. Randomisation will only succeed in breaking any correlation between the participants' preferences and the trial outcome if these former remain ignorant about the comparative nature of the experiment. But if they have strong preferences about the treatments, how far can we go in deceiving them about the comparison?

In order to control for such post-randomisation effects, Duflo et al. (2007) suggest two additional strategies. The first is continue collecting data after the experiment is terminated in order to verify whether the interaction with the experimenter was making any difference in the behaviour of the participants (e.g. Duflo and Hanna 2006). One way or another, we need participants to remain ignorant about the controls: they should not know they are still being observed. And we need to test this ignorance, just as in medical trials with blinding; we just cannot take it for granted.

To sum up, in RFEs, randomisation may generate a self-selection bias; we can only avoid with a partial or total masking of the allocation procedure. We have argued that this is a viable solution only insofar as the trial participants do not have strong preferences about the trial outcome. If they do, we cannot assume that blinded randomisation will be a control for their preferences unless we test for its success. We will only be able to claim that the trial has been impartial regarding the participants' preferences if we have a positive proof of them being ignorant of the comparative nature of the experiment. Hence, in RFEs, randomisation is not a strong warrant of impartiality per se: we need to prove in addition that it has been masked successfully.[1]

6 Can Field Trials Ground an Evidence-Based Policy?

In order to use RCTs as regulatory tools, it is necessary to provide some warrant of their external validity and impartiality. If we could have perfect causal knowledge of the effects of an intervention, impartiality would be warranted by default. But if there is uncertainty about it, RCTs should incorporate some warrants of impartiality.

[1] For a further discussion of the possibility of dispensing with randomisation in field experiments, see Deaton (2010) and Imbens (2010).

A regulatory decision should be impartial, and if we are going to ground it on inconclusive evidence, we need to make sure that nobody exploits such uncertainty in their own interest.

We have seen that the use of randomisation requires expert judgement, so the *mechanical objectivity* of RCTs is mere appearance: we need a subjective (judgement-based) assessment of its actual implementation in order to decide about the external validity of the trial. Nonetheless, randomisation provides a warrant of the impartiality of a clinical trial at the crucial stage of allocating treatments – not beyond that. Such warrant contributes to the credibility of the experimental outcome: we may question its external validity, but at least we can presume it is unbiased – at least more than unconstrained expert judgement.

As regulatory tools, RCTs have proven to be most successful at the FDA, where they are part of a system in four phases: the first two provide causal background knowledge for the trial, and the last one – post-marketing surveillance – controls for possible lacks of external validity. The number of label revisions, on the one hand, and market withdrawals, on the other, signals the levels of uncertainty with which the FDA is dealing. Randomisation, together with other means (such as blinding), has contributed to the impartiality of such uncertain regulatory decisions, making them more acceptable to the American public. However, it seems as though belief in the regulatory system has weakened today, and a debate has been started on how to strengthen the fairness of the FDA regulatory process.

If we are going to adopt RFEs as a public policy tool, we will probably need to work on two fronts. In Sect. 5, we have argued that randomisation needs to be successfully hidden from experimental subjects in order to be a warrant of impartiality in field trials, since we are not dealing with the biases of the researchers alone, but also with the preferences of the experimental subjects. On the other hand, as Cartwright has argued, and the example of the FDA seems to illustrate, a randomised trial per se does not warrant the external validity of its conclusions. We need to keep a record of the fallibility of the conclusions of field trials in order to measure the degree of uncertainty we are dealing with.

If we follow the institutional paradigm of the FDA, the question is how to integrate RFEs into an institutional system that makes their results credible. As of today, there is no clear answer as to which sort of institution should this be (Duflo and Kremer 2005). Government-sponsored programmes are rare because it is difficult to attain the high level of political consensus required for a successful implementation. Without this consensus, RFEs can easily fall prey to the sort of manipulations described in the previous section, in which each party will try to make the experiment support its views. Non-governmental organisations (NGOs) 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 biases: the culture of the organisation 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 programme.

NGOs (or non-profit organisations in general for that matter) have also a problem of credibility, not unlike the pharmaceutical industry: they usually have a stake in the programmes they evaluate (Pritchett 2002). And randomisation 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 organisations 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 organisations should provide the impartial expertise required to make the trials credible to the involved parties. This is probably the best solution. However, it remains an open question why would the participants in the trial see the international organisation as a neutral third party they can trust. Only if they do, one can be certain that the trials it sponsors are a credible source of knowledge about their target population.

Acknowledgements Our most sincere thanks to Hsiang-Ke Chao and Szu-Ting Chen for organising the very hospitable and intellectually fruitful conference in which this chapter was originally presented. Thanks to the editors and reviewers for their comments. Teira's research has been funded by the Spanish Ministry grant FFI2011-28835.

References

Akerlof, George. 1970. The market for 'Lemons': Quality uncertainty and the market mechanism. *Quarterly Journal of Economics* 84(3): 488–500.

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

Bolton, Gary E., Jordi Brandts, and Axel Ockenfels. 2005. Fair procedures: Evidence from games involving lotteries. *The Economic Journal* 115(506): 1054–1076.

Carnevale, Anthony Patrick, Stephen J. Rose, and Century Foundation. 2003. Socioeconomic status, race/ethnicity, and selective admissions. http://www.tcf.org/Publications/White%5FPapers/carnevale%5Frose.pdf. Accessed 15 Jan 2012.

Carpenter, Daniel P. 2010. *Reputation and power: Organizational image and pharmaceutical regulation at the FDA*. Princeton: Princeton University Press.

Carpenter, Daniel, and Colin Moore. 2007. 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.

Cartwright, Nancy. 2007. Are RCTs the gold standard? *BioSocieties* 2(1): 11–20.

Cartwright, Nancy, and Eileen Munro. 2010. The limitations of randomized controlled trials in predicting effectiveness. *Journal of Evaluation in Clinical Practice* 16(2): 260–266.

Deaton, Angus. 2010. Instruments, randomization, and learning about development. *Journal of Economic Literature* 48: 424–455.

Duflo, Esther. 2010. *La politique de l'autonomie*, Lutter contre la pauvreté, vol. 2. Paris: Seuil.

Duflo, Esther, and Michael Kremer. 2005. Use of randomization in the evaluation of development effectiveness. In *Evaluating development effectiveness*, World Bank Series on Evaluation and Development, vol. 7, ed. George Keith George, Osvaldo N. Feinstein, and Gregory K. Ingram, 205–232. New Brunswick/London: Transaction.
Duflo, Esther, and Rema Hanna. 2006. Monitoring works: Getting teachers to come to school: C.E. P.R. discussion papers, CEPR discussion papers: 5426. London
Duflo, Esther, Rachel Glennerster, and Michael Kremer. 2007. Using randomization in development economics research: A toolkit: C.E.P.R. discussion papers, CEPR discussion papers: 6059. London
Epstein, Steven. 1996. *Impure science. Aids and the politics of knowledge*. Berkeley: University of California Press.
Featherstone, Katie, and Jenny L. Donovan. 2002. "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(5): 709–719.
Hacking, Ian. 1988. Telepathy: Origins of randomization in experimental design. *Isis* 79(3): 427–451.
Heckman, James. 1992. Randomization and social policy evaluation. In *Evaluating welfare and training programs*, ed. F. Manski and Garfinkel Irwin, 201–230. Cambridge/London: Harvard University Press.
Imbens, G. 2010. Better LATE than nothing: Some comments on Deaton (2009) and Heckman and Urzua (2009). *Journal of Economic Literature* 48: 399–423.
Lasagna, L. 1959. Gripesmanship: A positive approach. *Journal of Chronic Diseases* 10: 459–468.
Macklin, Ruth. 2004. *Double standards in medical research in developing countries*, Cambridge Law, Medicine, and Ethics, vol. 2. Cambridge/New York: Cambridge University Press.
Marks, Harry M. 1997. *The progress of experiment. Science and therapeutic reform in the United States, 1900–1990*. New York: Cambridge University Press.
Marks, Harry M. 2000. Trust and mistrust in the marketplace: Statistics and clinical research, 1945–1960. *History of Science* 38: 343–355.
Orr, Larry L. 1999. *Social experiments: Evaluating public programs with experimental methods*. Thousand Oaks: Sage.
Petryna, Adriana. 2009. *When experiments travel: Clinical trials and the global search for human subjects*. Princeton: Princeton University Press.
Porter, Theodore M. 1995. *Trust in numbers: The pursuit of objectivity in science and public life*. Princeton: Princeton University Press.
Pritchett, Lant. 2002. It pays to be ignorant: A simple political economy of rigorous program evaluation. *Journal of Policy Reform* 5(4): 251–269.
Sackett, David, William Rosenberg, Muir Gray, Brian Haynes, and Scott Richardson. 1996. Evidence-based medicine: What it is and what it isn't. *British Medical Journal* 312: 71–72.
Stone, Peter. 2007. Why lotteries are just? *The Journal of Political Philosophy* 15(3): 276–295.
Teira, D. 2011a. Frequentist versus Bayesian clinical trials. In *Philosophy of medicine*, ed. Fred Gifford, 255–297. Amsterdam: Elsevier.
Teira, D. 2011b. *Impartiality in clinical trials*. London: University College London.
Urbach, Peter. 1985. Randomization and the design of experiments. *Philosophy of Science* 52(2): 256–273.
Wilson, Charles. 2008. Adverse selection. In *The New Palgrave dictionary of economics*, 2nd ed. Steven N. Durlauf and Lawrence E. Blume. London: Palgrave Macmillan.
Worrall, John. 2002. What evidence in evidence-based medicine? *Philosophy of Science* 69(3 Supplement): S316–S330.
Worrall, John. 2007. Why there's no cause to randomize. *The British Journal for the Philosophy of Science* 58(3): 451–488.

Chapter 12
Explaining the Explanations of 100 Million Missing Women

Hsiang-Ke Chao and Szu-Ting Chen

Abstract This chapter studies the methodology in the missing-women debate among economists and biologists. One of the central philosophical and methodological issues at stake in the missing-women debate is natural and social scientists' attempts for discovering the underlying causal structures and mechanisms. Although they encounter the same problem of inferring the mechanism and causal structure in face of available data, the discovering strategies vary. In this chapter, we will comparatively study the strategies of discovering causes and mechanisms in the case of missing women.

1 Amartya Sen's Missing Women

Nobel laureate economist Amartya Sen opened his 1990 *New York Review of Books* article, "More Than 100 Million Women Are Missing," with the following sentences:

> It is often said that women make up a majority of the world's population. They do not. This mistaken belief is based on generalizing from the contemporary situation in Europe and North America, where the ratio of women to men is typically around 1.05 or 1.06, or higher. In South Asia, West Asia, and China, the ratio of women to men can be as low as 0.94, or even lower, and it varies widely elsewhere in Asia, in Africa, and in Latin America. How can we understand and explain these differences, and react to them?

H.-K. Chao (✉)
Department of Economics, National Tsing Hua University,
101, Section 2, Kuang Fu Road, 30013 Hsinchu, Taiwan
e-mail: hkchao@mx.nthu.edu.tw

S.-T. Chen
Graduate Institute of Philosophy, National Tsing Hua University,
101, Section 2, Kuang Fu Road, 30013 Hsinchu, Taiwan
e-mail: stchen@mx.nthu.edu.tw

H.-K. Chao et al. (eds.), *Mechanism and Causality in Biology and Economics*, History, Philosophy and Theory of the Life Sciences 3, DOI 10.1007/978-94-007-2454-9_12,