PhD degree in Systems Medicine (curriculum in Foundations of life sciences, bio-ethics and cognitive sciences)

European School of Molecular Medicine (SEMM) and University of Milan

Settore disciplinare: M-FIL/02

CLINICAL TRIALS AND DRUG REGULATION: A PHILOSOPHICAL INQUIRY

Mattia Andreoletti

IEO, Milan

Matricola n. R10742

Supervisor: Prof. Maria Rescigno

IEO, Milan

Added Supervisor: Prof. David Teira

UNED, Madrid

Anno accademico 2016-2017

"6.53 The correct method in philosophy would really be the following: to say nothing except what can be said, i.e. propositions of natural science—i.e. something that has nothing to do with philosophy—and then, whenever someone else wanted to say something metaphysical, to demonstrate to him that he had failed to give a meaning to certain signs in his propositions. Although it would not be satisfying to the other person—he would not have the feeling that we were teaching him philosophy—this method would be the only strictly correct one" (Ludwig Wittgenstein, Tractatus Logico-Philosophicus).

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

21CCA = 21st Century Cures Act

EBM = Evidence-Based Medicine

FDA = Food and Drug Administration

RCT = Randomized Controlled Trial

GENERAL INTRODUCTION

"What is your aim in philosophy? -- To show the fly the way out of the fly-bottle" (Ludwig Wittgenstein, PI 309).

Philosophy of medicine: a personal stance

Until the late 70s, philosophy of science was still conceived as a theory of scientific knowledge, investigating the problem of its foundations, the problem of its methods, and the problem of the application of the resulting technologies, considering the aims of science also from an ethical perspective. In other words, there was a view of science as a unitary enterprise, ultimately reducible to physics. While, in more recent years, from the 80s onwards, philosophers of science have begun to recognize the different features of each scientific discipline: chemistry, biology, neurosciences, medicine, social sciences, and so on and so forth.

Within the *philosophy of special sciences*, we can identify two main approaches. The first one aims to discuss topical issues of philosophy of science – such as explanation, causation, confirmation, demarcation, etc. – within the relevant science, in order to make some philosophical progress, say, for instance, revising (again) the notion of scientific explanation or providing a new (yet another) normative theory of causation. Whereas, the second approach aims precisely at the opposite, directly addressing scientific issues, discussing them applying philosophical tools and reasoning in order to actively contribute to the scientific progress.

Between the two approaches, there is, first and foremost, a fundamental difference in the understanding of what exactly constitutes a philosophical problem. For the former, philosophical problems are essentially inherited from the past. Traditionally, there are topics which fall within the sole competence of philosophy: causation, for instance, has been discussed since the time of Aristotle. On the contrary, for the latter approach, the ones of tradition are not the only issues which deserve philosophical investigation, but there is much more worthy of philosophical analysis embedded in science. Broadly speaking, philosophical problems are those which people disagree about and for which empirical evidence cannot provide any certain answer. Then, there is little dispute that this sort of problems is common in every scientific discipline. Ethical issues are precisely that sort of problems, and they are the easiest to recognize, especially in biomedical sciences, because everyone is somehow familiar with ethics. By contrast, methodological and conceptual problems are usually a matter only for specialists. Therefore, it is hard to recognize them for everyone who has never been dealing with science.

By definition, empirical evidence cannot be brought to resolve methodological and conceptual problems. For example, which experimental design is better to implement in order to confirm a particular hypothesis is not something taken for granted nor is possible to get such a decision on the basis of an experiment. It is precisely in such areas that a philosophical approach could be helpful both for science and scientists. In such a way, there is hope of providing evidence that philosophy is not just a lot of hot air. As a matter of fact, philosophers are trained in tackling such methodological and conceptual questions, and they have skills which better enable them to identify those problems. We firmly believe that a philosopher who has been also trained in a scientific discipline can even settle the issues of that discipline definitively. With the appropriate scientific training, his approach can help to clarify

the problems, enlighten the potential solutions, eliminate some unreasonable options, provide convincing arguments, and avoid inconsistencies. From this point of view, philosophy of science is a growing and evolving discipline with an increasing impact on society.

Among the philosophies of special sciences, *philosophy of medicine* is quite an emerging field. Albeit the relationship between philosophy and medicine dates back to ancient time, it has emerged as an academic discipline only recently. One of the fundamental and most long-standing debates in the philosophy of medicine relates to the basic concepts of health and disease (see Boorse 1975). Exploring this distinction remains epistemologically and morally important as these definitions influence when and where people seek medical treatment, and whether society regards them as "ill", including whether they are permitted or urged to receive medical treatments. The dividing line between disease and health is notoriously vague, and there is a great deal of disagreement in the literature in philosophy of medicine.

In the 80s the privileged epistemological status of medicine has been questioned from various perspectives: social anthropologists and historians have explored its cultural contingency; medical sociologists described medicine "as an institution of social control and the locus of professional power struggles for cognitive authority and control" (Richards 1988); influential and eminent figures within and outside medicine - such as Cochrane (1972) and Illich (1981) - have criticized its autonomy and its efficacy. As well, philosophers considering medicine in the context of science emphasized its distinctiveness (or inferiority) compared to physical sciences.

The Evidence-Based Medicine movement that arose in the 90s can be seen partly as an effort to make medicine more scientific, grounding the "practice" of medicine in theoretical more methodologically robust disciplines, such as clinical epidemiology. In general, the impetus for EBM can be attributed to an increasing

awareness of the weaknesses of standard clinical practices and their impact on both the quality and costs of healthcare in the United States. The EBM movement mostly focused on how we should assess the safety and efficacy of medical therapies, in a way that can lead clinical decisions. According to an overused definition/quotation:

> *"Evidence-based medicine is the conscientious, explicit, and judicious use of current best evidence in making decisions about the care of individual patients" (Sackett et al. 1996).*

The main concern of EBM advocates was that too often physicians relied merely on their judgment and experience to make a clinical decision, which over the years have been proved to be misleading. To provide an alternative and a definition of what counts as "best evidence", proponents of EBM have developed a "hierarchy of evidence" that categorize different research methods with respect to their supposed quality. At the top of the hierarchy of evidence are Randomized Controlled Trials (RCTs). Thus, evidence produced by RCTs has been called the "gold standard" of evidence in EBM. "However, to refer to RCTs as the "gold standard" of evidence suggests that they are more. Specifically, one may be led to assume that RCTs are necessary for reliable causal inference or that RCTs are guaranteed to deliver reliable results" (Reiss and Ankeny 2016). That is why, in the last decade, many philosophers of medicine have scrutinized RCTs in all their facets.

Briefly, RCTs are tightly controlled experiments for the evaluation of interventions[1]. The major difference between them and other experimental designs is the random allocation of the participants in the two group of the trial. Randomization

[1] Note that a RCT is not merely a clinical trial, but they can also be employed in many other contexts: for instance, to evaluate economic or socio-educational interventions. Since we are here in a medical context, I limit myself to mentioning clinical RCTs.

is a key feature because it protects against selection bias[2], it provides the basis for statistical inference (Armitage 1982, 2003), and it permits masking. Nancy Cartwright and John Worrall more than anyone else challenged the first two assumptions. In particular, they hold that randomization is neither essential for statistical tests nor it can control for all confounders (Cartwright 2010, 2011, 2012; Cartwright and Hardie 2015; Worrall 2007, 2010a, 2010b). Moreover, RCTs are often criticized for their low external validity. The distinction between internal versus external validity of clinical trials refers to how well a study answers certain questions. In general, a clinical trial has high internal validity "if the results can be considered accurate for the sample included in the study" (La Caze 2016). Whereas, external validity "is the degree to which the results of an observation hold true in other setting" (Fletcher, Fletcher, and Wagner 1996). While the internal validity depends on the capacity of an experimental design to control for biases and confounders, the external validity requires something more, such as an understanding of how the intervention under testing works (causal knowledge) and of the context in which the intervention is to be employed. Therefore, as suggested by many scholars (e.g. Howick, Glasziou, and Aronson 2013; Clarke et al. 2013; Illari 2016), we must rely on other sources of evidence, for instance mechanistic knowledge or experts' judgment, in order to improve the external validity of a randomized trial.

Because of these epistemic limitations, RCTs appear to be increasingly ill-adapted to influence medical decision-making, therefore philosophers of science have recently argued for an epistemic pluralism. Nobody suggests that researchers should give up RCTs. Instead, they urge the supplementation of RCTs with other forms of evidence: we should avoid reduction to a single method. There are indeed different

[2] However, it is curious, as Mebius suggests, that "there has been no clear evidence to date to support the claim […] that randomization reduces selection bias" (Mebius 2014).

forms of medical knowledge that must be integrated with RCTs to make a causal inference more robust. For instance, many philosophers of sciences (e.g. Clarke et al. 2013; Campaner 2011; Illari 2011; Illari 2016; Russo and Williamson 2007) have emphasized the importance of mechanisms, which are held to be essential for causal inference. As of today, three research projects are leading the way in the field: EBM+ (Clarke et al. 2013), Cause Health, and Philpharm (Landes, Osimani, and Poellinger 2017). All the projects aim at defending a pluralistic approach to causal inference in medicine.

The debate surrounding the foundations of EBM makes the nature of philosophical disputes in medicine even more evident: they are not only theoretically interesting *per se* but also because they might have an impact on medical practice. Insofar as we are concerned more with the practical consequences of this theoretical debate, regulatory issues come along (see Reiss 2010).

Clinical trials and drug regulation: a philosophical inquiry

This thesis examines some recent controversies surrounding the evaluation of medical treatments and the organization of drug regulation. A significant issue is whether or not the current approach to the regulation of new medicines should be modified to manage issues generated by new biomedical products. One might reasonably question whether regulatory systems with their roots in the early part of the twentieth century are actually capable of dealing with the issues and problems posed by the molecular revolution.

As mentioned above, many philosophers of science have already scrutinized the epistemological features of RCTs, mostly undermining their status as the "gold standard" method for assessing causality (e.g. Cartwright 2010; Russo and Williamson 2007; Worrall 2007). While, critics of pharmaceutical system (Gøtzsche 2013; Goldacre

2014; Ioannidis 2016) believe that epistemological limitations of clinical trials are a backdoor for promoting the interests of pharmaceutical companies. This thesis aims at broadening the philosophical analysis of clinical trials beyond epistemology to re-appraise their epistemic import within the socio-political context in which they are embedded. Hopefully, this philosophical analysis would allow a better understanding of current debates on regulation of new medical products, facilitating the integration of microsocial level and epistemological level of analysis of regulatory standards of evidence.

These are exciting times for translational medicine as the convergence between fundamental and clinical research comes of age. As Boniolo and Nathan put it (Boniolo and Nathan 2016), "Molecular medicine is likely to become one of the next exciting frontiers of philosophical research", and just recently philosophers have started to investigate the philosophical implications of the molecular revolution in biology and medicine. Over the last decades, many scientific breakthroughs have enabled the development of increasingly complex medical treatments. Also, advances in biomedical sciences and technology have boosted the production of new medicines. Clearly, there has been an impressive acceleration after the first releases of the Human Genome Project, which has offered an actionable entry into virtually all diseases with a genetic basis. As an example, we have really achieved some major advancements in gene and stem cell therapy[3]. With the former, we can treat rare and less rare diseases, such as Wiskott-Aldrich syndrome (Aiuti et al. 2013) and hemophilia B (Nathwani et al. 2014), while stem cell therapy is now a commercial reality for corneal regeneration (Rama et al. 2010). Nonetheless, many are arguing that we did reach nothing more than that: biomedical research has not actually delivered as much as expected (e.g. Joyner,

[3] Gene therapy can be defined as the use of genetic material (usually DNA) to manipulate a patient's cells for the treatment of a disease. Cell therapy can be defined as the infusion or transplantation of whole cells into a patient.

Paneth, and Ioannidis 2016). In some areas, such as oncology, many medical needs are still unmet, and we are not seeing the breakthrough therapies that we expect given the enormous cognitive and financial resources put into basic research. On the one hand, our comprehension of the biological basis of disease is exponentially increasing; on the other hand, very few medicines reach the bed of the patients, and when it happens they often fail to deliver any real benefit on overall survival or quality of life (see Davis et al. 2017). This gap between basic and clinical research is widely known and it is usually described as the "valley of death" (Butler 2008). As the pharmaceutical industry productivity crisis worsens, there are calls for regulatory changes to support innovation.

Especially in cancer research, things are changing at a hectic pace. We cannot enumerate here the myriads of cancer therapies that have been developed in the last decade, nonetheless we are bound to mention the pivotal paper by Hanahan and Weinberg "The Hallmarks of Cancer" (Hanahan and Weinberg 2011). According to Hanahan and Weinberg, targeted therapies can be categorized according to their respective effects on hallmark capabilities of cancer. What is worth to highlight in the present context is that targeted therapies are highly selective, meaning that they act on specific molecular targets. This sort of treatments was simply unthinkable one century ago when the regulatory system emerged. This latter was designed indeed to provide massive consumer protection at a point when our understanding of the biology of cancer was still relatively poor and statistical tests gave the only solid evidence about treatment effects. Whereas, the majority of scientific breakthroughs in biomedicine (from cracking the genetic code to the discovery of restriction enzymes) took place after the regulatory reforms and would not bear fruit until the 70s and 80s. As of today, the exponential growth of treatments both in number and complexity is posing many challenges to regulators.

The basic idea underlying medical regulation entails that governments have an obligation and responsibility to protect the public from unreasonable harms. It is precisely for this reason that national regulatory agencies have been established in Western democracies. Of course, at the same time, a rational regulation must not stifle medical innovation, biomedical research, and technological development. Even under ideal circumstances, this balance is hard to reach. With this regard, the current gold standard to assess drug safety and efficacy (RCTs) is showing its age. On the one hand, reformists of the regulatory system claim that the current cumbersome standard is preventing many potentially lifesaving treatments from reaching patients' bed. On the other hand, critics are complaining about the many shortcomings of RCTs, which would provide low-quality scientific evidence, hardly answering any relevant clinical question, thus not offering any real protection to patients. Nonetheless, they both agree that the current critical trial system is broken[4], and a more rational regulatory system is desirable.

Even a superficial look at regulatory documents is sufficient to claim that in its core parts drug regulation has not evolved much over the years, especially in setting up scientific standards to assess safety and efficacy of innovative treatments. Although targeted therapy agents are increasingly available for clinical applications, many of these promising drugs have produced disappointing results when tested in clinical trials, indicating that there are many challenges that must be addressed to advance this field (Schilsky et al. 2010; Wistuba et al. 2011). Most standard trial designs are not optimal for testing targeted drug. For instance, in oncology one major issue is patient recruitment. Many therapies are targeting narrow populations of patients harboring a

[4] See for instance the speech that Janet Woodcock, director of FDA's Center for Drug Evaluation and Research, delivered at a recent workshop on real world evidence (RWE) at the National Academies of Sciences, Engineering, and Medicine (http://www.raps.org/Regulatory-Focus/News/2017/09/20/28500/FDAs-Woodcock-The-Clinical-Trials-System-is-Broken/ Accessed October 12, 2017).

specific genetic mutation, which are numbered in hundreds whereas standard trial designs require thousands of participants. Secondly, many of these drugs are cytostatic rather than cytotoxic, and thus they would not meet the usual end points when tested in a typical registration trial. While for standard chemotherapy the conventional method of measuring effectiveness is through tumor shrinkage, for target therapies time to progression or progression-free survival (PFS) better capture the beneficial effects. However, these outcomes are not considered as "objective" endpoints, and are not sufficient for regular market approval.

The practical challenges to meet the old standard of evidence on the one hand, and the pressures from companies and patients' associations on the other hand, pushed the regulators to grant more and more exceptions. Indeed, as a reaction to these pressure, the FDA has developed 4 different alternative regulatory pathways: fast track, breakthrough therapy, accelerated approval, priority review. On their own, clinical research and pharmaceutical companies have started to investigate alternative trial designs (e.g. adaptive trials). However, this continuous derogation from well-conducted RCTs has raised many concerns about the "quality" of regulatory decisions. These concerns have been motivated also by some notorious failures, as, for instance, in the case of Avastin®. Very briefly, Avastin® (bevacizumab), is an inhibitor of angiogenesis, targeting VEGF growth factor. The rationale of the drug is simple, since tumors need to develop blood vessels to grow and survive, the inhibition of angiogenesis could be a very effective therapeutic line. In 2008, Avastin has been tested in three different clinical trials for three different types of tumors, colorectal, lung and breast cancer, with positive results. However, subsequent attempts to replicate those findings in larger trials failed (see D'Agostino 2011; Sekeres 2011).

In general, the difficult question regulators are facing is when the study design and gathered data are good enough to be relied upon. For instance, while we should be

justifiably cautious about claiming that effects observed in small trials are causal (Pereira, Horwitz, and Ioannidis 2012), sometimes an effect is just too strong to be dismissed because there may be biases, but determining exactly when is a matter of contentious. Moreover, many breakthrough therapies are tested only with very rudimental experimental designs. The recent development and approval of chimeric antigen receptor T-cell therapy (Tisagenlecleucel/Kymriah®) for the treatment of refractory B-cell acute lymphoblastic leukemia (ALL) is a nice illustration thereof. CAR-T cell treatment starts from a person's own cells, isolates them from the body, engineers them to express a tumor-specific chimeric receptor (CAR), and puts them back in the body where they can attack cancer cells. The FDA approval of Kymriah® is based on the results of an open-label, multicenter, single-arm Phase II small trial (N=63). The FDA's decision raised many concerns, also because of the choice of the drug manufacturer (Novartis) to charge it 425,000 dollars. According to many scholars, evidence from different sources can be considered reliable in some circumstances, but there is not yet a clear consensus on that. In the case of Kymriah® the suspicion that emotional stories, such as that of Emily Whitehead (see (Rosenbaum 2017), have been a major drive of the approval decision, is legitimate. Thus, in such a case a call for more robust evidence may be reasonable, even though RCTs are hardly conceivable: patients recruitment and masking being the major issues, in addition to the enormous costs involved.

In conclusion, identifying issues with current evidential comparisons and evaluations of medical interventions, and indicating how they can be improved is philosophically interesting and it also has the potential to improve the scientific basis upon which regulatory systems are based. A rational approach to drug regulation should aim not only at making the evaluation of new products more efficient, but also at making these products available to patients more quickly and thereby to enhance

public health. Nonetheless, it remains to be elucidated whether new regulatory standards of evidence, being they either methodologically naive or statistically sophisticated, are reliable enough to guide sound regulatory decisions, and this inquiry is not reducible only to epistemic consideration. There is indeed one major philosophical claim in the subtext of this dissertation, who serves also as a general thread: epistemic considerations alone are not sufficient to capture the current controversies over regulatory standards of evidence. The question on how we should regulate medical treatments does not really resolve into a mere debate on causality.

In the following chapters, we will extensively explore some of the topics mentioned in this introduction. Shifting scientific landscape poses new challenges and requires the freedom and flexibility to continuously adapt and evolve. This also applies to philosophers who are tracking those landscapes, that is why we decided to branch away from the proto-book dissertation model, and organized this thesis as a collection of articles. Each chapter has been indeed conceived as a self-standing paper. To facilitate reading and examination, references are provided at the end of each chapter. The present short overview of regulatory issues is clearly not comprehensive but rather an attempt to give an idea about the complexity of this important area of work that has many direct links with philosophy of science.

Summary of chapters

Why do we need Randomized Controlled Trials?

In this chapter, we dig into the history of the RCTs in order to bring out and make clear the reasons why they became the gold standard for drug testing and regulation. In doing so, we focus on the evolution of drug regulation in the United States arguing that scandals played a crucial role in the development of drug regulations, forcing regulators to acknowledge the weak points of previous standards and to consider more

robust alternatives, ranging from laboratory test to clinical trials. Changes and reforms were implemented in the regulatory system as a response to major pharmaceutical scandals, and not in response to the real epistemic needs put in place by developments in drugs research.

The philosophical debate has focused on the methodological virtues and vices of RCTs, praising or blaming them to the extent that they could succeed in assessing causality. This "oversimplification of trials epistemology" has already been pointed out, but still, the role of scandals in fostering methodological progress in drug regulation is quite neglected. If we are right in claiming that regulators adopted RCTs in order to prevent pharmaceutical scandals, then we should evaluate the epistemic import of trial designs also to the extent to which they can prevent scandals.

More than one way to measure: a casuistic approach to cancer clinical trials

Nowadays RCTs are usually cumbersome experiments, expensive and costly. Clinical researchers and regulators have been almost blind to alternatives, focusing instead on large trials (large populations for statistical power), hypothesis testing, and control for type 1 error. While this is perfectly fine for many treatments, it is not satisfactory for new oncological ones. In the last years, science and technology made great progress towards a better understanding of fundamental biological mechanisms of the diseases. As of today, we know that each tumor has a different genomic basis despite its site of occurrence and is controlled by the local microenvironment. This genomic heterogeneity and complexity make the treatment of the disease nearly impossible with most of the current chemotherapies. In this chapter, we show that the gold standard for testing and approving drugs is not suitable to test new cancer treatments anymore and that it might be a bottleneck for progress in managing cancer, as argued by reformists of regulatory system. Then we claim that: (1) in some cases we have an

epistemological justification to abandon the obsession with the null-hypothesis testing, and consequently (2) there is no way by which FDA could still retain the old "one size fits all" approach. There is no doubt that the traditional two-armed randomized controlled trial revolutionized medicine, moving medical research from art to science, but now RCTs are likely the weakest link in the chain of scientific knowledge.

Rules versus standards: a legal-philosophical framework for drug regulation

Over the last decade, philosophers of science have extensively criticized the epistemic superiority of RCTs for testing safety and efficacy of new drugs, defending instead various forms of evidential pluralism. We argue that scientific methods in regulatory decision making cannot be assessed in epistemic terms only: there are various costs involved. Drawing on the legal distinction between rules and standards, we show that drug regulation based on evidential pluralism has much higher costs than our current RCT-based system. We analyze these costs and advocate for evaluating any scheme for drug regulatory tests in terms of concrete empirical benchmarks, like the error rates of regulatory decisions.

Drug regulation and evidentiary pluralism

In this chapter, we want to argue that the multiplicity of testing standards is more defensible than critics think. As a matter of fact, since 1962 we already have multiple testing standards for testing the safety and the efficacy of medical treatments, and the system has worked reasonably well so far. As we argue, medical treatments (not just drugs) have different testing standards according to the potential public health risks they pose. With regard to this, we present a concept of risk that, in our view, captures our current regulatory consensus. Risks depend on two factors: the hazards involved in a treatment and the number of people potentially exposed to it. From a political standpoint, this concept of risk is all we need to justify the existence of multiple testing

standards. Consequently, different testing standards are defensible for certain class of drugs (e.g. targeted therapies). Finally, we address a potential crucial objection to our argument: that any a priori assessment of risks always comes with an unacceptable degree of uncertainty, therefore we should always demand for stricter testing standards.

Statistical evidence and the reliability of medical research

In this chapter, we focus on the reliability of RCTs conducted to test the safety and efficacy of medical treatments. RCTs are scientific experiments and, as such, we expect them to be replicable. However, for more than a decade now we have been discussing a replicability crisis across different experimental disciplines including medicine: the outcomes of trials published in very prestigious journals often disappear when the experiment is repeated. First, we see how replicability and statistical significance are connected: we can only make sense of the p-value of a trial outcome within a series of replications of the test. But, in order to conduct these replications properly, we need to agree on the proper design of the experiment we are going to repeat. Then, we argue that trialists need to agree on the debiasing procedures and the statistical quality controls that feature in the trial protocol if they want the outcome to be replicable. Furthermore, we make two complementary points. On the one hand, replicability *per se* is not everything: we need trial outcomes that are not only statistically significant but also clinically relevant. On the other hand, trials are not everything: the experts analyzing the evidence can improve the reliability of statistical evidence, although they sometimes fail; we need to study further how they make their decisions.

Disclaimer

The following chapters, accordingly amended, have been published/submitted, or are in preparation for submission, as follows:

Andreoletti M. – Why do we need Randomized Controlled Trials? (in preparation)

Andreoletti M. – More than one way to measure? A casuistic approach to cancer clinical trials (Submitted to *Perspective in Biology and Medicine*)

Andreoletti M. & Teira D. – Rules versus Standards: a legal-philosophical framework for drug regulation (Submitted to *Science, Technology, and Human Values*)

Andreoletti M. – Drug regulation and evidentiary pluralism (in preparation)

Andreoletti M. & Teira D. – Statistical evidence and the reliability of biomedical research (in book: eds. Solomon M., Simon JR., Kincaid H., *The Routledge Companion to Philosophy of Medicine*, Routledge, 2016)

References

Aiuti, Alessandro, Luca Biasco, Samantha Scaramuzza, Francesca Ferrua, Maria Pia Cicalese, Cristina Baricordi, Francesca Dionisio, et al. 2013. "Lentiviral Hematopoietic Stem Cell Gene Therapy in Patients with Wiskott-Aldrich Syndrome." *Science* 341 (6148):1233151. https://doi.org/10.1126/science.1233151.

Armitage, Peter. 1982. "The Role of Randomization in Clinical Trials." *Statistics in Medicine* 1 (4):345–52. https://doi.org/10.1002/sim.4780010412.

———. 2003. "Fisher, Bradford Hill, and Randomization." *International Journal of Epidemiology* 32 (6):925–28. https://doi.org/10.1093/ije/dyg286.

Boniolo, Giovanni, and Marco J. Nathan. 2016. *Philosophy of Molecular Medicine: Foundational Issues in Research and Practice*. Taylor & Francis.

Boorse, Christopher. 1975. "On the Distinction between Disease and Illness." *Philosophy & Public Affairs* 5 (1):49–68.

Butler, Declan. 2008. “Translational Research: Crossing the Valley of Death.” *Nature News* 453 (7197):840–42. https://doi.org/10.1038/453840a.

Campaner, Raffaella. 2011. “Understanding Mechanisms in the Health Sciences.” *Theoretical Medicine and Bioethics* 32 (1):5–17. https://doi.org/10.1007/s11017-010-9166-5.

Cartwright, Nancy. 2010. “What Are Randomised Controlled Trials Good For?” *Philosophical Studies* 147 (1):59. https://doi.org/10.1007/s11098-009-9450-2.

———. 2011. “A Philosopher’s View of the Long Road from RCTs to Effectiveness.” *The Lancet* 377 (9775):1400–1401. https://doi.org/10.1016/S0140-6736(11)60563-1.

———. 2012. “RCTs, Evidence, and Predicting Policy Effectiveness,” August. https://doi.org/10.1093/oxfordhb/9780195392753.013.0013.

Cartwright, Nancy, and Jeremy Hardy. 2012. *Evidence-Based Policy: A Practical Guide to Doing It Better*. New York, USA: Oxford University Press USA. http://global.oup.com/?cc=gb.

Clarke, Brendan, Donald Gillies, Phyllis Illari, Federica Russo, and Jon Williamson. 2013. “The Evidence That Evidence-Based Medicine Omits.” *Preventive Medicine* 57 (6):745–47. https://doi.org/10.1016/j.ypmed.2012.10.020.

D’Agostino, Ralph B. Sr. 2011. “Changing End Points in Breast-Cancer Drug Approval — The Avastin Story.” *New England Journal of Medicine* 365 (2):e2. https://doi.org/10.1056/NEJMp1106984.

Davis, Courtney, Huseyin Naci, Evrim Gurpinar, Elita Poplavska, Ashlyn Pinto, and Ajay Aggarwal. 2017. “Availability of Evidence of Benefits on Overall Survival and Quality of Life of Cancer Drugs Approved by European Medicines Agency: Retrospective Cohort Study of Drug Approvals 2009-13.” *BMJ* 359 (October):j4530. https://doi.org/10.1136/bmj.j4530.

Fletcher, Robert H., Suzanne W. Fletcher, and Grant S. Fletcher. 2012. *Clinical Epidemiology: The Essentials*. Lippincott Williams & Wilkins.

Goldacre, Ben. 2012. *Bad Pharma: How Drug Companies Mislead Doctors and Harm Patients*. London: Fourth Estate.

Gøtzsche, Peter C. 2013. *Deadly Medicines and Organised Crime: How Big Pharma Has Corrupted Healthcare*. Radcliffe Publishing.

Hanahan, Douglas, and Robert A. Weinberg. 2011. "Hallmarks of Cancer: The Next Generation." *Cell* 144 (5):646–74. https://doi.org/10.1016/j.cell.2011.02.013.

Howick, J, P Glasziou, and Jk Aronson. 2013. "Can Understanding Mechanisms Solve the Problem of Extrapolating from Study to Target Populations (the Problem of 'External Validity')?" *Journal of the Royal Society of Medicine* 106 (3):81–86. https://doi.org/10.1177/0141076813476498.

Illari, Phyllis. 2016. *Mechanisms in Medicine*. Routledge Handbooks Online. https://doi.org/10.4324/9781315720739.ch5.

Illari, Phyllis McKay. 2011. "Mechanistic Evidence: Disambiguating the Russo–Williamson Thesis." *International Studies in the Philosophy of Science* 25 (2):139–57. https://doi.org/10.1080/02698595.2011.574856.

Ioannidis, John P. A. 2014. "More Than a Billion People Taking Statins?: Potential Implications of the New Cardiovascular Guidelines." *JAMA* 311 (5):463–64. https://doi.org/10.1001/jama.2013.284657.

———. 2016. "Evidence-Based Medicine Has Been Hijacked: A Report to David Sackett." *Journal of Clinical Epidemiology* 73 (Supplement C):82–86. https://doi.org/10.1016/j.jclinepi.2016.02.012.

Joyner, Michael J., Nigel Paneth, and John P. A. Ioannidis. 2016. "What Happens When Underperforming Big Ideas in Research Become Entrenched?" *JAMA* 316 (13):1355–56. https://doi.org/10.1001/jama.2016.11076.

La Caze, Adam. 2016. *The Randomized Controlled Trial: Internal and External Validity*. Routledge Handbooks Online. https://doi.org/10.4324/9781315720739.ch18.

Landes, Jürgen, Barbara Osimani, and Roland Poellinger. 2017. "Epistemology of Causal Inference in Pharmacology." *European Journal for Philosophy of Science*, March, 1–47. https://doi.org/10.1007/s13194-017-0169-1.

Mebius, Alexander. 2014. "Corroborating Evidence-Based Medicine." *Journal of Evaluation in Clinical Practice* 20 (6):915–20. https://doi.org/10.1111/jep.12129.

Nathwani, Amit C., Ulreke M. Reiss, Edward G.D. Tuddenham, Cecilia Rosales, Pratima Chowdary, Jenny McIntosh, Marco Della Peruta, et al. 2014. "Long-Term Safety and Efficacy of Factor IX Gene Therapy in Hemophilia B." *New England Journal of Medicine* 371 (21):1994–2004. https://doi.org/10.1056/NEJMoa1407309.

Pereira, Tiago V., Ralph I. Horwitz, and John P. A. Ioannidis. 2012. "Empirical Evaluation of Very Large Treatment Effects of Medical Interventions." *JAMA* 308 (16):1676–84. https://doi.org/10.1001/jama.2012.13444.

Rama, Paolo, Stanislav Matuska, Giorgio Paganoni, Alessandra Spinelli, Michele De Luca, and Graziella Pellegrini. 2010. "Limbal Stem-Cell Therapy and Long-Term Corneal Regeneration." *New England Journal of Medicine* 363 (2):147–55. https://doi.org/10.1056/NEJMoa0905955.

Reiss, Julian. 2010. "In Favour of a Millian Proposal to Reform Biomedical Research." *Synthese* 177 (3):427–47. https://doi.org/10.1007/s11229-010-9790-7.

Reiss, Julian, and Rachel A. Ankeny. 2016. "Philosophy of Medicine," June. https://plato.stanford.edu/archives/sum2016/entries/medicine/.

Richards, Evelleen. 1988. "The Politics of Therapeutic Evaluation: The Vitamin C and Cancer Controversy." *Social Studies of Science* 18 (4):653–701. https://doi.org/10.1177/030631288018004004.

Rosenbaum, Lisa. 2017. "Tragedy, Perseverance, and Chance — The Story of CAR-T Therapy." *New England Journal of Medicine* 377 (14):1313–15. https://doi.org/10.1056/NEJMp1711886.

Russo, Federica, and Jon Williamson. 2007. "Interpreting Causality in the Health Sciences." *International Studies in the Philosophy of Science* 21 (2):157–70. https://doi.org/10.1080/02698590701498084.

Sackett, David L., William M. C. Rosenberg, J. A. Muir Gray, R. Brian Haynes, and W. Scott Richardson. 1996. "Evidence Based Medicine: What It Is and What It Isn't." *BMJ* 312 (7023):71–72. https://doi.org/10.1136/bmj.312.7023.71.

Schilsky, Richard L., Jeff Allen, Joshua Benner, Ellen Sigal, and Mark McClellan. 2010. "Commentary: Tackling the Challenges of Developing Targeted Therapies for Cancer." *The Oncologist* 15 (5):484–87. https://doi.org/10.1634/theoncologist.2010-0079.

Sekeres, Mikkael A. 2011. “The Avastin Story.” *New England Journal of Medicine* 365 (15):1454–55. https://doi.org/10.1056/NEJMc1109550.

Wistuba, Ignacio I., Juri G. Gelovani, Jörg J. Jacoby, Suzanne E. Davis, and Roy S. Herbst. 2011. “Methodological and Practical Challenges for Personalized Cancer Therapies.” *Nature Reviews Clinical Oncology* 8 (3):nrclinonc.2011.2. https://doi.org/10.1038/nrclinonc.2011.2.

Worrall, John. 2007. “Why There’s No Cause to Randomize.” *The British Journal for the Philosophy of Science* 58 (3):451–88. https://doi.org/10.1093/bjps/axm024.

———. 2010a. “Do We Need Some Large, Simple Randomized Trials in Medicine?” In *EPSA Philosophical Issues in the Sciences*, 289–301. Springer, Dordrecht. https://doi.org/10.1007/978-90-481-3252-2_27.

———. 2010b. “Evidence: Philosophy of Science Meets Medicine.” *Journal of Evaluation in Clinical Practice* 16 (2):356–62. https://doi.org/10.1111/j.1365-2753.2010.01400.x.

1. WHY DO WE NEED RANDOMIZED CONTROLLED TRIALS?

"What has history to do with me? Mine is the first and only world" (Ludwig Wittgenstein, NB p. 82).

1.1 Introduction

As all the medical-scientific community knows, Randomized Control Trials (RCTs) are the main experimental study design to evaluate the efficacy of a specific treatment in a given population. Conventionally, the term "treatment" refers to many kinds of interventions: diagnostic, screening, health education, etc. Although RCTs are systematically and extensively adopted in the drug research and testing, as they are the last phase of a mandatory threefold process, which is strictly regulated by transnational laws. Of course, RCTs did not come out of the blue, nor did the rules that had made them compulsory. In this chapter, we dig into the history of the randomized controlled trials in order to bring out and make clear the reasons why they became the gold standard for drug testing and regulation. In doing so, we focus on the evolution of drug regulation in the United States (Gaudillière & Hess 2012; Marks 1997; Temin 1980, 1985). In particular, we argue that pharmaceutical scandals played a crucial role in the development of drug regulations, forcing regulators to acknowledge the weak points of previous standards and to consider more robust alternatives, ranging from laboratory tests to RCTs. The historical investigation of the evolution of methodological concepts would be sufficient to warrant our claim (Schickore 2011). When, why and in

what context did regulators implement RCTs as the gold standard for drugs testing? As we argue, changes and reforms were implemented in response to major pharmaceutical scandals, and not in response to the real epistemic needs put in place by developments in drugs research. On the contrary, these epistemic needs remained largely unsatisfied.

Our claim is in contrast with the mainstream epistemology of clinical trials, which consider them as the result of a methodological development and scientific progress of medicine. Many philosophers of science and physicians have argued indeed that RCTs have been designed to assess a genuine casual relation between the drug and its effects. Then, the debate has focused on the methodological virtues and vices of RCTs, praising or blaming them to the extent that they could succeed in assessing causality (Cartwright 2010, 2011; Clarke et al. 2013; Howick 2011; Worrall 2010b, 2010a). This "oversimplification of trials epistemology" has already been pointed out (Hey 2015), but still the role of scandals in fostering innovations in drug regulation is quite neglected among philosophers of science. If we are right in claiming that regulators adopted RCTs in order to avoid pharmaceutical scandals, then one should evaluate the epistemic import of trial designs also to the extent to which they could prevent scandals. Taking into account historical and socio-political context is particularly relevant for the recent debate on the adoption on new regulatory standards (Avorn and Kesselheim 2015). As the historian of medicine, Marcia Meldrum put it: "the RCT is a dynamic methodology, and its present and future are informed by its history" (Meldrum 2000).

Historically, the link between scandals and policies in Western democracies is nothing new: many sociologists and political scientists have discussed it for decades (Butler, Drakeford, and Butler 2005; Thompson 2013). In general, a scandal is defined as an event, often regarded as morally wrong, which causes public outrage. While is

clear that scandals play a crucial role in the general political scenario, is quite uncharted whether these events could have an impact in other fields[1], such as clinical research. In the following paragraphs we are going to show that they had a crucial role in triggering reforms in drug regulation, and hence in shaping the methodology of contemporary clinical research. RCTs served at best the goals of regulators.

1.2 The great American fraud

In the last decades of the nineteenth century, laboratory science had a great boost thanks to the development of many basic research fields such as chemistry, physiology, and microbiology. These scientific advancements ended up in what historian of medicine Charles Rosenberg (Rosenberg 1997) has called a "therapeutic revolution", that is, the discovery of a noticeable number of effective therapeutic agents. Physicians and patients were deeply affected by this "revolution", as they came across a continuously increasing number of new drugs.

Private companies manufactured most of the drugs that were placed on the market in those years: their share in the U.S. drug market was the 72% and their business was continuously growing. At that time, the chemical composition of almost all the compounds was kept strictly secret to protect intellectual property and patent. However, physicians realized soon that many drugs did not contain any active ingredient, but pharmaceutical companies promoted the inactive drugs in the same way as the ones with real and active compounds. This is why, for instance, Samuel Hopkins Adams, an American investigative journalist (a *muckraker*), in 1905 coined the expression "The Great American Fraud" referring to the drug trade situation. In discussing therapeutic reforms, the market plays a contingent yet significant role, as

[1] Carpenter (Carpenter 2010) and Porter (Porter 1996) are partial remarkable exceptions. While Hutchinson (Hutchison 2016) has recently made a similar point, but focusing on nursing practice.

much as scientific progress does. Indeed, at a certain point, the medical scientific community had to face "a novel intellectual and political problem" (Marks 1997) how to foster even further the increasing scientific progress in the laboratories while protecting the patients and the market from fake and potentially unsafe drugs. In other words, there was the need to tell apart effective and ineffective drugs without discrediting the entire scientific enterprise.

The American Medical Association (AMA) made the first effort towards a more rational approach to pharmaceutical therapeutics. However, while the *Journal* of the association was reviewing and publishing the best results of medical research, it also printed advertisements for some very low-quality drugs. In order to solve this apparent contradiction, on the spring of 1905 the AMA established the Council on Pharmacy and Chemistry, which had the task of investigating the medicines advertised in the pages of the *Journal*. Very soon, the Council thought that its scope had to be extended to all new medicines available to physicians and not only to those advertised in the JAMA. The work of the Council was to review the scientific evidence supporting a drug and deliberate on its quality. In practice, the scientific evidence was often scarce and then the deliberation of the council reflected "a curious mixture of judgments […] and opinions" (Marks 1997). When the council's assessment was a matter of laboratory tests, in order to reveal whether the drug contained an active known ingredient, the decision was quite easy. However, pharmaceutical companies developed also drugs containing ingredients that could be tested in a laboratory, but whose beneficial properties were completely unknown. In these murky cases, the deliberation was more difficult or even impossible. In these latter cases, extra-scientific considerations, such as the track record of the companies, played a major role in the deliberation process. As just mentioned, clinical evidence was often scarce or even missing; the Council relied on the expertise of academic clinicals but often bumped into opinions too much

different. Hence, they warned that their approval for the biological purity of the compounds did not imply clinical efficacy. In many cases, laboratory tests could not address the question of efficacy. Take for instance glandular extracts (e.g. red bone marrow, ovarian, parotid gland extracts) that were common on the market in the early 1900s. Usually, labels reported correctly the chemical composition, and this could be easily tested in the laboratories. However, it was unclear what all those extracts actually did: laboratory tests were not sufficient for that question.

Nonetheless, the U.S. Government in 1906 passed a first key-legislation to control drug market: the Pure Food and Drug Act. The new law gave to the Bureau of Chemistry (the predecessor to the FDA) in the Department of Agriculture the legal power to seize adulterated or misbranded products (Junod 2008). But it assumed the same standards of the Council: laboratory tests to check whether a drug contained the ingredients labeled or advertised by the manufacturer. However, the law did not allow anyone to screen drugs and control for potential frauds before their placing on the market: it was remedial but not preventive. Moreover, the meaning and the exact enforcement of the 1906 Act were questionable. In 1912, to counter this flaw, the U.S. Congress enacted the Sherley Amendment that prohibited explicitly false therapeutic claims. However, in the following years, the consequences of the new law were practically nil, since it was still hard to prove something regarding the therapeutic effects of the drug through laboratory tests. The necessity to investigate in a more systematic way a method to test for drug efficacy was made clearer by some emerging scandals.

One of the most striking was the case of Banbar, an old patent medicine advertised by the producer as a cure for diabetes. The drug was not dangerous *per se,* since it contained just inactive ingredients like milk, sugar and a grass plant known as "equisetum". Nonetheless, it was obviously life threating for those who rejected insulin,

which had become a standard treatment short after its discovery in 1922 and whose effectiveness was beyond any dispute or doubt. Meanwhile, in 1927, the Bureau of Chemistry's name was transformed into Food, Drug, and Insecticide Administration, then abbreviated to the current version (FDA). In the 30s the "new" FDA accused the producer of Banbar of fraud and took to court all the evidence about the death of patients who had refused to take insulin in order to get Banbar. Conversely, in its defense, the producer of the drug took to the court testimonial letters, which consumers had written thanking him. Those letters were sufficient to demonstrate to the court his *bona fides* about the efficacy of the drug. Thus, the FDA did not get the authorization to seize the product, so it remained on the market (Junod 2008).

These were the clear limits of the 1906 Act: it was more about basic chemical quality control (the drug actually had the ingredients it claimed it had) in order to protect consumers from frauds, and therefore preventing potential scandals, rather than addressing more relevant epistemic needs such as *safety* and *efficacy*. These would come in the following decades, when new scandals made it unavoidable.

1.3 The 1938 Food, Drug, and Cosmetic Act

In the wake of Banbar and other minor scandals, people started being more and more suspicious of pharmaceutical companies and drug trade. In those years, two books became very popular and influential among the public opinion: *100,000,000 Guinea Pigs: Dangers in Everyday Foods, Drugs, and Cosmetics* by Arthur Kallet and F.J Schilnk (Kallet and Schlink 1932) and *American Chambers of Horrors: the truth about food and drugs* by Ruth deForest Lamb (deForest Lamb and Copeland 1936). The authors harshly criticized the FDA and the government for their failure in protecting people from the abuses and the frauds of drug companies. In particular, they pointed out all

the weakness of the 1906 Act, asking for an immediate update. Instead, at the very beginning the FDA reacted vindicating the success of all its activities.

In the 1930s, more than a hundred companies were manufacturing drugs containing sulfanilamide (Marks 1997), a "wonder" antibacterial compound used to cure streptococcal infections. The company S. E. Massengill decided to produce syrup-type sulfanilamide using diethylene glycol, an extremely toxic solvent. The syrup was placed on the market without any tests in animals or humans, causing at least 106 documented deaths (Wax 1995). However, under the 1906 Act, the FDA could only prosecute Massengil for misbranding. The subsequent public outrage prompted the Congress to pass a new set of laws: the 1938 *Food, Drug and Cosmetic Act*. The 1938 Act required companies to inform the FDA of their intention to put a new drug on the market. On the one hand, the FDA was given the power to ask for "adequate tests by all methods reasonably applicable to show whether or not the drug is safe". The major concern of regulators in 1938 Act was the safety of the drugs, whereas they did not nearly consider the problem of evaluating the efficacy, which of course they soon bumped into. On the other hand, the 1938 Act did not make FDA approval a pre-requisite for market access (Marks 1995).

Let us focus on the kind of "adequate tests" required by the FDA as proof of the drug safety. Although these tests remained unspecified in the act, the regulators adopted the same standards already advocated by the AMA's Council on Pharmacy and Chemistry: laboratory analysis and experts' evaluation. Moreover, animal tests, even if not formally required, were systematically requested by the FDA and became soon a sort of gold standard for drug safety. This was one of the major novelties of the 1938 Act. Another major accomplishment was the overcoming of the "fraud flaw" of 1906 Act: the FDA could now remove from the market unsafe drugs without having to prove that there was intent of fraud on the part of the producer.

Already in 1938, the new Act was put to the test. In the spring of 1938 British researchers had discovered a new sulfonamide compound (2-para-aminobenzene pyridine), apparently better than every other sulfa drug. Experimental tests in mice showed low toxicity, few adverse side effects, and more beneficial effects than its predecessors. In October 1938, Merck & Company, an American company, submitted an application for the FDA to approve *sulfapyridine* for the treatment of pneumonia, for which there was no effective therapy yet. The FDA requested the opinion of the experts and clinicians who had had the opportunity to test the experimental drug. Some of them were reporting adverse events, some did not. On the drug's efficacy, the data were even more unconvincing: the drug had been administered only to a few patients with pneumonia and it was still too early to judge its efficacy. This is why many skeptics were advising FDA to keep the application on hold since they were concerned about the risk-benefit balance. They were also concerned about the lack of data on the effects of sulfapyridine on other infectious diseases for which it might be prescribed.

The FDA had adopted the view that the expert judgment of qualified clinical investigators should prevail over the opinion of regular clinicians. But in case of disagreement among the former, the debate would not be settled by the methodological superiority of their respective tests, but through the majority rule. Despite the pressure of the press, asking for a fast approval of the drug, and despite the incoming winter, a time when cases of pneumonia were obviously more frequent, the FDA kept collecting and reviewing data and experts' opinions until the deadline provided for in the statute. On March 1939 the FDA decided to "not deny" (NB: officially, the act did not allow the FDA to approve a drug, but just gave to the agency the power to deny a request) the applications for sulfapyridine, provided that manufacturers explicitly reported on the labels and in advertising that the drug had to be used "under close, continuous observation of a qualified practitioner of medicine"

(Marks 1997). This is because some doubts remained about the efficacy of the drug as noted by Theodore Klumpp, by then chief of the Drug Division in the FDA: "While a few investigators recommended that the drug be withheld from the market such recommendations upon analysis do not appear to rest upon considerations of the intrinsic safety or danger of the drug. Principally those workers were concerned with the orderly development of medical scientific knowledge, concerning the therapeutic efficacy of the drug […]" (Marks 1997). The sulfapyridine was soon replaced by a more powerful drug, penicillin, so the extent of the FDA's decision is not clear. But at least regarding the safety "adequate tests", laboratory analysis and experts' judgment, gave the impression to perform that task well. At least, it seemed so.

What was clear among the medical community, at that point, was that the standards adopted by the FDA was far from being able to check for efficacy. Drug evaluation basically was left to the judgments and opinions of experts, which was considered superior to regular clinical judgment, and medical community thought to be reliable at least in spotting adverse effects. However, another scandal would soon undermine that belief and forced the FDA to reconsider again its regulations and standards. Developments outside the medical field, in statistics applied to agriculture to be exact, converged to make it possible. In the next paragraph we are going to briefly present major ones.

1.4 Statistics: today's innovations for tomorrow's standards

Physicians had been dealing with the variability of biological phenomena for centuries. They were always aware of the fundamental role of chance in medical observations: the natural course of the disease, spontaneous remissions, and responses to treatments were considerably different in each patient. Clinical measurement was not as uniform as laboratory tests. Therefore, physicians relied only on their experience in order to

handle uncertainty. This was the case also in comparative experiments. Indeed, knowledge of the variance of the diseases and of potential perturbing factors could be exploited then to perform comparative studies, trying to reduce the chance to a minimum. Of course, managing chance was considered fundamental for any comparative experiment. Therefore, their quality depended on the experience of the researcher. Still this approach had serious limitations because physicians' knowledge of both confounding factors and the magnitude of natural random variability might be limited. What statistics could offer to clinical researchers was an experimental design that permitted to manage biological variability and chance regardless of previous knowledge. Generally, this breakthrough is credited to the genius of a British statistician and biologist, Sir Roland Aylmer Fisher (1890-1962).

Fisher had been dealing with biological variability since 1919 when he began to work as statistician at the agricultural experimental station in Rothamsted. In fact, Fisher had to find a reliable method to solve some practical problems in agricultural research: Which varieties of seeds are better? Which fertilizer? Which crop rotation system is best? Simple comparisons cannot provide a reliable answer. Suppose that you observe a 10 percent difference in yields between two grain varieties: is it due to a real difference in the quality of the seeds or to plot conditions? One way to answer this is to rely on experience: an expert farmer could tell that a 10 percent difference is never due to plot conditions alone. Nonetheless for Fisher, this strategy was far from being scientific since it relied entirely on experts' knowledge (i.e. subjective). Moreover, it would not be feasible if such previous knowledge were not available to anyone. Another option would be to replicate the experience many times, but this is rarely possible in agricultural practice. Indeed, Fisher calculated that it would require approximately five hundred years to find that such a 10 percent difference is due to chance alone, Fisher's solution consisted in setting up a new experimental design. He

divided the experimental plots in strips in order to increase the number of observations in a single experiment. This way he reduced the variability of the effects due to other factors than a quality difference between grains. In other words, he increased the sample size of the experiment. But the most crucial innovation was to sow grain in strips in a random order. According to Fisher, randomization is the only mechanism that could ensure the validity of scientific inference in a comparative experiment. The randomization of the plots ensured that all the possible perturbing factors were equally distributed among all the strips. In Fisher's own words, randomization helped to protect the experimenter from a devilish nature.

According to Fisher randomization is essential not just for controlling for confounders, but also for the calculation of the probability of finding a given difference between the experimental treatments. This idea was illustrated by the famous thought experiment of the lady testing tea. Suppose, says Fisher, that a lady declares that she can tell whether milk or tea was poured first to a cup, just by tasting. What kind of experiment can one design to test her assertion? According to Fisher, it would consist in preparing 4 cups of tea pouring milk first and 4 pouring milk later and present them to the lady in a random order, determined by "the apparatus used in games of chance" (Fisher 1937). She has to spot the 4 cups prepared pouring milk first (or tea first). The random order in presenting her the cups is important, according to Fisher, because it guarantees that the probability of guessing rightly all the cups by chance is 1 out of 70. If the order is not random and the lady spots it in some way, the chance of guessing the cups right by chance increases and this can bias the result of the experiment. How many cups has she to guess right in order to prove her wonder tasting? If she were right on every cup, of course, one should obviously accept it. But what about 3 rights and 1 wrong for instance? One should set a threshold of "significance", that is setting a degree of probability to refute what Fisher called the null hypothesis, our default assumption.

In this case, that the lady cannot tell whether it had been milk or tea. Conventionally this threshold is fixed at 5% (the standard level of statistical significance)[2] and one should ignore all the results which fail to reach this standard. The probability of guessing 3 cups right and 1 wrong by chance is 16 out of 70, which is more than 20%. Thus, one should not discard the null hypothesis. One would be willing to do it if and only if the lady guesses right 4 cups out 4, for which the chance goes beyond 5%.

As noted by Marks (Marks 1997), Fisher's direct influence on biological and medical communities was negligible. It was Bradford Hill, a British statistician working on medical topics, who exported Fisher's experimental design to drug testing in the 1940s. Historians of medical statistics have argued, time and again, that British physicians did not grasp the statistical rationale of randomization (Armitage 1982; Chalmers 2011). There was instead a widespread concern among British doctors about the many ways in which personal biases could spoil the evaluation of novel therapies. They found in the randomized allocation of treatment a device that could neutralize the personal beliefs of investigators as to who would benefit most from the therapy. Allocation bias occurs when the allocation of subjects to study groups is jeopardized by the preferences of the experimenters (e.g. the healthiest or youngest patients receive the experimental treatment). Randomization can easily succeed in neutralizing this bias. However, there are many other biases which can occur in a comparative experiment. For instance, participants' preferences can still spoil the result, conditioning the evaluation of the outcomes. If physicians want to favor the drug under testing, they could report better outcomes for the experimental drug and so could do patients as well. That is why we need another de-biasing method, such as blinding the allocation of treatments to physicians and patients. Randomization is an essential part

[2] On the origin of this conventional level of statistical significance see for instance (Cowles and Davis 1982)

of blinding procedures, as regarded for Fisher. However, to be precise, it is not randomization *per se*, which guarantees the fairness of trials, but the experimental design as a whole, which includes basic statistical analysis such as test of significance and controls. "The use of properly designed clinical trials permits us to move from an authoritative frame of reference to a scientific one" (Marks 1997).

Nevertheless, it took a decade to implement Fisher's approach in medicine[3]: the first randomized controlled trial, with significance testing, took place in Britain in 1947. It would take one more decade to spread among physicians and two more decades to transform it into a regulatory standard[4]. As it will be clear in the following section, it is always difficult to change the minds of scientists without empirical evidence, and to change the minds of politicians and regulators without scandals and public outrage.

1.5 How randomized controlled trials became the gold standard

In the years after the war, some major breakthroughs in clinical trials design were achieved in two independent studies of streptomycin. For the first time, researchers introduced in trials' design a standardized set of controls that will become soon fundamental: a control group, the random allocation of patients, and standardized non-qualitative criteria to assess outcome. In the U.S., the Public Health Serviced (PHS) organized a research study on streptomycin to treat tuberculosis. PHS researchers did not want to make the mistakes of their predecessors, so they strictly controlled the trial in order to get a reliable knowledge about the use of the drug. Moreover, the scarce

[3] Actually, RCTs were initially resisted also in agriculture, and the competitor method – the half-drill strip – remained in wide use for a long time beyond the arrival of RCTs. This is because of their complexity and because it was nearly impossible for the farmers to understand it (see Berry, 2015).

[4] In the 1958 Donald Mainland, a medical statistician, attending a meeting of Endocrinology Society in San Francisco, noted a "statistical attidute" among the panelists. New concepts and methods were circulating, whereas just a decade before "anyone who advocated them was commonly regarded as an aberrant specimen" (Mainland 1960).

funding and limited amount of streptomycin available made it necessary to arrange comparative experiments in order to produce the best knowledge in the most efficient way. In order to control for the allocation bias, the study design included the randomization of treatments. PHS researchers' main concern was to avoid individual decisions of physicians, especially those who were already convinced of the beneficial effect of streptomycin. That is also why PHS researchers planned to conduct the entire study in a double-blind fashion, but they failed to convince the involved physicians. Nonetheless, the study produced reliable and uncontested results in favor of streptomycin. However, it employed only descriptive statistics, there was no use of statistical tests of significance.

On the other side of the Atlantic, in 1947, the British Medical Research Council was conducting a very similar trial, which became known as the "first RCT" since it employed for the first time a standardized method for statistical inference. The scientist in charge was Sir Austin Bradford Hill, a relevant actor in the history of medicine. Indeed, he contributed much to the methodology of clinical trials, publishing a series of papers on medical statistics in *The Lancet* journal (1937) claiming the relevance of randomization and controls to ensure the objectivity of a study. In particular, Bradford Hill argued that the primary experimenter's aim is "to ensure beforehand that, as far as possible, the control and treated groups are the same in all relevant respects" (Yoshioka 1998). Moreover, randomization was crucial to ensure the objective assessment of treatments since it removed personal responsibility from the clinician from selecting which patients would benefit. These two ideas shaped the rationale behind the design of MRC trial. The trial enrolled 107 patients randomized in two groups: 55 assigned to the experimental group receiving streptomycin and the standard of care (bed rest) and 52 to the control group receiving only bed rest. The radiologists who interpreted x-ray chest exams were blind to the allocation of the

treatments. After 6 months there were only 4 deaths in the streptomycin group, whereas there were 15 in the control. Investigators considered that difference statistically significant, "the probability of it occurring by chance is less than one in a hundred" (Marshall et al. 1948). For the first time both a method for minimize allocation bias and statistical evaluation of collected data were employed in a clinical trial. That is why Hill's trial became (very slowly) a milestone and influenced an entire generation of physicians, even though it was not without opposition.

Certainly, these trials had both a great and important weight in the history of medicine and clinical research, i.e. the exclusion of subjective judgments from drugs testing and evaluation. Rather than relying on conflicting opinions of individual physicians based on different standards, the new methodological standard provided a more objective and scientific tool to appraise therapeutic innovations. However, that standard was integrated into drug regulation more than a decade later, in the aftermath of further pharmaceutical scandals.

The Fifties (and later the Sixties) were the golden age of antibiotics: more than 400 drugs were introduced into the market each year (Meldrum 2000). Though, most of them were nothing but "me too drugs", "which were defined as molecular alterations that did not show evidence of therapeutic superiority over the pioneer or, in the case of fixed-combination drugs, over the component drugs used alone" (Carpenter 2010). Pharmaceutical companies invested much in advertising, reporting published studies showing beneficial effects, but most of those reports were of questionable validity. Usually, pharmaceutical companies sent (directly or through a physician) some sample of their new drugs to the doctors for a try, asking them to report their experiences. Of course, this practice was completely unsatisfactory from a scientific point of view, and yet it was sufficient for receiving the FDA's approval. As we mentioned above, the old

1938 statute did not provide methodological standards for clinical research, and its power to control access to market was limited.

In 1958, the U.S. Senator Estes Kefauver held hearings on the drug industry. His main concern was the exorbitant profit margins of pharmaceutical companies, due mostly to antibiotics. The companies blamed it on the high costs of research since that many drugs failed during drug development. The hearings generated important evidence documenting the poor quality of clinical research supporting the marketing of many drugs. It revealed to the public what all the experts already knew: most of the clinical research was just rubbish.

Another tragedy triggered the enactment of the Kefauver-Harris 1962 Amendments of the 1938 Act and the subsequent Investigational New Drug Regulations in 1963. Kefauver's hearing placed drug regulation on the top of the agenda of U.S politics. The story of thalidomide is well-known. As Daniel Carpenter has pointed out, the standard narrative comes from an American journalist, Morton Mintz, who published an article about thalidomide in the Sunday morning Washington Post on July 15, 1962. According to Carpenter "Mintz essentially re-interpreted bureaucratic nitpicking and deliberation […] as modern-day, scientific virtues that upheld protection of American families and infants" (Carpenter 2010). Thalidomide was a quite popular drug in Europe and especially in Western Germany, where the drug was manufactured by pharmaceutical company Chemie Grünenthal since 1957 and marketed as Contergan. The drug was prescribed to treat a great number of various symptoms, mostly psychological as anxiety or tension. But it was also administered to many pregnant women to alleviate nausea and sickness. This was the beginning of a tragedy for thousands of women around the world. Those who had taken thalidomide gave birth to children with phocomelia, a terrible congenital disorder involving limbs malformations, leading to premature death. In the U.S, the German company reached

an agreement with Richardson-Merrell to market the drug, and this latter filed the application for approval with the FDA in 1961 when evidence of thalidomide side effects started to be reported. Both the German and the American companies denied the link between the cases of phocomelia and its product. As part of the approval process, the drug was distributed to many physicians in the U.S for testing purposes. But, at the FDA, one of the physicians reviewing thalidomide approval, Dr. Frances Oldham Kelsey, decided to withhold it asking for more clinical tests, because evidence of serious adverse effects was already appearing in Germany and in other 20 countries where the drug had been approved. Kelsey's decision was indeed a great and fortunate one and it has secured her a place in history. Unfortunately, the testing samples still caused 17 reported cases of phocomelia. Under public pressure and after a rush discussion, the Congress passed in 1962 a new pharmaceutical regulatory framework, inspired by Kelsey's precautionary attitude. First, it introduced a system of control by FDA over clinical experimentation, assigning an IND (Investigational New Drug) status to experimental drugs, and nullifying this status if clinical trial protocols were not methodologically sound or patients' rights were not respected. Second, it removed the "automatic" approval by default after 60 days: drugs needed a "positive" approval by the FDA to enter the market. And third, above all, it required "substantive evidence" of effectiveness based on "well-controlled studies", in addition to the pre-clinical demonstration of safety. The lawmakers left the task of better specifying the meaning of those expressions to FDA experts and officers, who saw the minimum standard in "randomized controlled trials". Moreover, the 1963 IND rules shaped somehow the 3-phases structure of drug testing, the form DF 1571 listed for the first time three phases of trial:

"a. Clinical pharmacology. This is ordinarily divided in two phases: Phase 1 starts when the new drug is first introduced into men

[…]; phase 2 covers the initial trials on a limited number of patients for specific disease control […]. b. Clinical trial. This phase 3 provides the assessment of the drug's safety and effectiveness […]. A reasonable protocol is developed on the basis of the facts accumulated in the earlier phases, including completed and submitted animal studies". This phase is conducted by separate groups following the same protocol […] to produce well-controlled clinical data" (Carpenter 2010).

Let us focus briefly on this division, since it provides additional evidence to our claim that RCTs become the gold standard in medical research because they better served the political goals of regulators, compared to animal experiments and experts judgment. Indeed, it is the design of phase III trials that makes possibile to objectivly assess the safety and efficacy of a drug, all the previous phases are pointless to this epistemic aim. However, as it was already clear at that time, RCTs were quite challenging and demanding experiments requiring many patients in order to allow correct statistical inferences - "the method of controlled trials is still in its infancy; that, although the principles are simple, the art is extremely difficult" (Mainland 1960). Intuitevly, running big experiments in humans can rise a medical scandal as well, exposing many individuals to a drug that can be potentially toxic. This possibility would result in an even bigger scandal than thalidomide, making the fears of early critics of pharmaceutical industry (Kallet & Schlink 1932) true: actually turning people into guinea pigs[5]. Therefore, the early phases were introduced in order to provide preliminary evidence of safety, before exposing many patients to the drug. From a

[5] Some years later (1972) the infamous Tuskgee study of syphillis (1972) led to further major changes in U.S. law and regulation on the protection of participants in clinical studies. It is no coincidence that recent medical scandals and public concerns over clinical research have been triggered mostly by accidents during the experimental phases (see, Eddleston, Cohen, & Webb 2016), rather than post-market failuress (e.g. drugs withdrawals).

purely epistemic point of view these phases are negligible, but from a political point of view they served to protect consumers from medical disasters.

We have provided some historical evidence to warrant the idea that a series of pharmaceutical scandals pushed drug regulation in the US in the direction of tighter and tighter controls, leading ultimately to the adoption of RCTs as the current safety and efficacy standard. In conclusion, RCTs have been adopted in response to the pressure of pharmaceutical consumers in Western democracies through parliaments.

1.6 Conclusion

In this chapter, we have tried to provide a straightforward answer to the question "Why do we need randomized controlled trials?", without engaging controversial epistemological features, such as causality assessment. To this end, we have briefly traced the history of drug regulation in order to understand why randomized controlled trials have been adopted as the gold standard for drug testing. To sum up, we have shown that at the beginning of the last century the drug trade grew and expanded quickly. Soon, emerging scandals made clear that some producers were selling fake drugs, making remarkable profits. Then, quality controls were adopted as the first standard for drug regulation: chemical laboratory tests became necessary in order to get market approval. In the following years, developments of new drugs led to some safety issues. Laboratory tests alone could not deal with safety of compounds, for which instead a very different standard was necessary, and animal pre-clinical research was still quite unreliable. Regulators relied then on experts' judgment and their clinical experience in order to control safety. However, very soon another issue came at stake: it was not possible to assess the safety of the drug regardless of its efficacy, and expert's deliberation was inadequate to account for it. Often it was impossible to reach a consensus among clinicians, especially about the efficacy of a

drug. At the same time, innovations in the design of experiments and statistics emerged outside the medical field, and gave the opportunity to the medical community to conceive and conduct methodologically sound comparative experiments to test innovative treatments, i.e. RCTs. Despite some early successful applications of the new experimental design in the context of medical testing, FDA had neither the political nor the scientific authority to set new standards for clinical research. Indeed, the FDA held its previous standards until another tragedy occurred in 1962. Thalidomide scandal forced regulators to revise their previous decisions, and to finally implement randomized controlled trials as the gold standard for drug testing. Once again, a scandal was crucial in triggering a reform of drug regulation.

References

Armitage, Peter. 1982. "The Role of Randomization in Clinical Trials." *Statistics in Medicine* 1 (4):345–52. https://doi.org/10.1002/sim.4780010412.

———. 2003. "Fisher, Bradford Hill, and Randomization." *International Journal of Epidemiology* 32 (6):925–28. https://doi.org/10.1093/ije/dyg286.

Avorn, Jerry, and Aaron S. Kesselheim. 2015. "The 21st Century Cures Act — Will It Take Us Back in Time?" *New England Journal of Medicine* 372 (26):2473–75. https://doi.org/10.1056/NEJMp1506964.

Berry, Dominic. 2015. "The Resisted Rise of Randomisation in Experimental Design: British Agricultural Science, c.1910–1930." *History and Philosophy of the Life Sciences* 37 (3):242–60. https://doi.org/10.1007/s40656-015-0076-8.

Butler, Ian, and Mark Drakeford. 2003. *Social Policy, Social Welfare and Scandal*. London: Palgrave Macmillan UK. https://doi.org/10.1057/9780230554467.

Carpenter, Daniel. 2014. *Reputation and Power: Organizational Image and Pharmaceutical Regulation at the FDA*. Princeton University Press.

Cartwright, Nancy. 2010. “What Are Randomised Controlled Trials Good For?” *Philosophical Studies* 147 (1):59. https://doi.org/10.1007/s11098-009-9450-2.

———. 2011. “A Philosopher’s View of the Long Road from RCTs to Effectiveness.” *The Lancet* 377 (9775):1400–1401. https://doi.org/10.1016/S0140-6736(11)60563-1.

Chalmers, Iain. 2011. “Why the 1948 MRC Trial of Streptomycin Used Treatment Allocation Based on Random Numbers.” *Journal of the Royal Society of Medicine* 104 (9):383–86. https://doi.org/10.1258/jrsm.2011.11k023.

Clarke, Brendan, Donald Gillies, Phyllis Illari, Federica Russo, and Jon Williamson. 2013. “The Evidence That Evidence-Based Medicine Omits.” *Preventive Medicine* 57 (6):745–47. https://doi.org/10.1016/j.ypmed.2012.10.020.

Cowles, Michael, and Caroline Davis. 1982. “On the Origins of the .05 Level of Statistical Significance.” *American Psychologist* 37 (5):553–58. https://doi.org/10.1037/0003-066X.37.5.553.

Eddleston, Michael, Adam F. Cohen, and David J. Webb. 2016. “Implications of the BIA-102474-101 Study for Review of First-into-Human Clinical Trials.” *British Journal of Clinical Pharmacology* 81 (4):582–86. https://doi.org/10.1111/bcp.12920.

Fisher, Ronald Aylmer. 1937. *The Design of Experiments*. Oliver And Boyd; Edinburgh; London.

Gaudillière, Jean-Paul, and V. Hess. 2012. *Ways of Regulating Drugs in the 19th and 20th Centuries*. Springer.

Group, British Medical Journal Publishing. 1948. “Streptomycin Treatment of Pulmonary Tuberculosis: A Medical Research Council Investigation.” *Br Med J* 2 (4582):769–82. https://doi.org/10.1136/bmj.2.4582.769.

Hey, Spencer Phillips. 2015. “What Theories Are Tested in Clinical Trials?” *Philosophy of Science* 82 (5):1318–29. https://doi.org/10.1086/683816.

Howick, Jeremy. 2011. *The Philosophy of Evidence-Based Medicine*. Oxford, UK: Wiley-Blackwell. https://doi.org/10.1002/9781444342673.

Hutchison, Jacqueline S. 2016. “Scandals in Health-Care: Their Impact on Health Policy and Nursing.” *Nursing Inquiry* 23 (1):32–41. https://doi.org/10.1111/nin.12115.

Junod, Suzanne W. 2008. FDA and Clinical Drug Trials: A Short History," in *A Quick Guide to Clinical Trials*, Madhu Davies and Faiz Kerimani, eds. (Washington: Bioplan, Inc.: 2008), pp. 25-55.

Kallet, Arthur, and Frederick John Schlink. 1932. *100,000,000 Guinea Pigs: Dangers in Everyday Foods, Drugs, and Cosmetics*. Vanguard Press.

Lamb, Ruth deForest. 1936. *American Chamber of Horrors: The Truth About Food and Drugs*. New York, Farrar. http://archive.org/details/americanchambero00lamb.

Mainland, Donald. 1960. "The Clinical Trial—Some Difficulties and Suggestions." *Journal of Chronic Diseases* 11 (5):484–96. https://doi.org/10.1016/0021-9681(60)90013-8.

Marks, H M. 1995. "Revisiting 'the Origins of Compulsory Drug Prescriptions'." *American Journal of Public Health* 85 (1):109–15. https://doi.org/10.2105/AJPH.85.1.109.

Marks, Harry M. 1997. *The Progress of Experiment: Science and Therapeutic Reform in the United States, 1900-1990*. Cambridge History of Medicine. Cambridge [England] ; New York: Cambridge University Press.

Meldrum, Marcia L. 2000. "A BRIEF HISTORY OF THE RANDOMIZED CONTROLLED TRIAL: From Oranges and Lemons to the Gold Standard." *Hematology/Oncology Clinics of North America* 14 (4):745–60. https://doi.org/10.1016/S0889-8588(05)70309-9.

Porter, Theodore M. 1996. *Trust in Numbers: The Pursuit of Objectivity in Science and Public Life*. Princeton University Press.

Rosenberg, Charles E. 1997. *No Other Gods: On Science and American Social Thought*. JHU Press.

Schickore, Jutta. 2011. "What Does History Matter to Philosophy of Science? The Concept of Replication and the Methodology of Experiments." *Journal of the Philosophy of History* 5 (3):513–32. https://doi.org/10.1163/187226311X599934.

Temin, Peter. 1980. *Taking Your Medicine: Drug Regulation in the United States*. Harvard University Press.

———. 1985. “Government Actions In Times Of Crisis: Lessons From The History Of Drug Regulation.” *Journal of Social History* 18 (3):433–38. https://doi.org/10.1353/jsh/18.3.433.

Thompson, John B. 2013. *Political Scandal: Power and Visability in the Media Age*. John Wiley & Sons.

Wax, Paul M. 1995. “Elixirs, Diluents, and the Passage of the 1938 Federal Food, Drug and Cosmetic Act.” *Annals of Internal Medicine* 122 (6):456. https://doi.org/10.7326/0003-4819-122-6-199503150-00009.

Worrall, John. 2010a. “Do We Need Some Large, Simple Randomized Trials in Medicine?” In *EPSA Philosophical Issues in the Sciences*, 289–301. Springer, Dordrecht. https://doi.org/10.1007/978-90-481-3252-2_27.

———. 2010b. “Evidence: Philosophy of Science Meets Medicine.” *Journal of Evaluation in Clinical Practice* 16 (2):356–62. https://doi.org/10.1111/j.1365-2753.2010.01400.x.

Yoshioka, Alan. 1998. “Use of Randomisation in the Medical Research Council’s Clinical Trial of Streptomycin in Pulmonary Tuberculosis in the 1940s.” *BMJ* 317 (7167):1220–23. https://doi.org/10.1136/bmj.317.7167.1220.

2. MORE THAN ONE WAY TO MEASURE: A CASUISTIC APPROACH TO CANCER CLINICAL TRIALS

"Philosophers constantly see the method of science before their eyes, and are irresistibly tempted to ask and answer questions in the way science does. This tendency is the real source of metaphysics, and leads the philosopher into complete darkness" (Ludwig Wittgenstein, Blue and Brown Books).

2.1 Introduction

Over the last years, science and technology have made great progress towards a better understanding of fundamental biological mechanisms of the diseases. Physicians, relying just on their own clinical experience, have long recognized that each patient is different from every other patient in many aspects. It is a matter of simple facts that many patients died without responding to any treatment, while others with the same (supposed) disease perfectly survived[1]. Notably, in oncology the variability of treatment response has been a long-standing problem. Nowadays, thanks to genomics ad post-genomics advancements we are now finding out an indication for that individual variability: tumor *heterogeneity*. Now, we know that each tumor has a different genomic basis despite its site of occurrence. This means that a lung cancer, for example, can have a molecular profile more similar to a melanoma than to another

[1] First evidence of individual variability dates back to the 6th century BC, when Pythagoras noted that ingestion of fava beans resulted in a potentially fatal reaction in some, but not all, individuals (Pirmohamed 2001).

lung cancer. In addition, every type of human cancer is comprised of subsets of different cell populations, with a different genetic background, meaning that, in theory, every cancer patient, from a genomic standpoint, might have an orphan disease.

On the one hand, the genomic complexity of cancer makes the treatment of the disease highly inefficient with most of the current chemotherapies. Indeed, tumor heterogeneity is a key determinant of tumor progression and a leading cause of failure of current anti-cancer treatments and emergence of drug-resistance (Diaz et al. 2012): tumors harboring different aberrations respond differently to the same drugs. On the other hand, the better the understanding of tumor heterogeneity the higher the chance to exploit it for treatments, and this is going to revolutionize the way we understand medicine.

Also, from a more theoretical point of view, tumor heterogeneity is posing many epistemological challenges that have not been discussed extensively by philosophers of science, with some noteworthy exceptions (Bertolaso 2011; Boniolo and Nathan 2016; Germain 2012; Plutynski 2013). For instance, Boniolo (Boniolo and Nathan 2016) argues that tumor heterogeneity is a major drive of methodological revolution of molecular medicine, that is the practice of medicine informed by the discoveries of molecular biology. Historically, clinical and experimental methods have been conceived as diametrically opposed to each other: the former dealing with the individual/particular, the latter dealing with the general/universal. Nowadays the two methods somehow fuse together in a sort of "hybrid method totally unthinkable" in the past (Boniolo and Nathan 2016). This novelty is epitomized by the introduction into molecular biology laboratories of new experimental models: the so-called *primary tumor cell cultures*, that are cells directly derived from a patient's primary tumor tissue sample; the patient-derived tumor xenografts, immunodeficient mice implanted with patients' tumors, and IPSc, reprogrammed stem cells from patients' fibroblasts which

may be further differentiated into cell lines which otherwise would be particularly difficult to obtain, such as neurons. All these cutting-edge methods aim at creating some sort of avatars of human patients in order *to bring the clinic into the lab*, allowing the researchers to study the individual with the methods of experiments and controls which are typical of the "science of universal".

Unravelling the basis of individual variability and hijacking cancer genomes seems the right strategy to win the "war on cancer". However, the methodological revolution described by Boniolo has not yet had a counterpart in clinical research, and this represents a concrete bottleneck for therapeutic progress (Richard Simon 2010). In this chapter, we question the resistance to change of the current drug regulatory paradigm. We present our case in the following order. At first, we discuss the implications of tumor heterogeneity for regulatory assessment of new medications, considering the challenges that the current paradigm is facing. Secondly, we take into account some criticisms that could explain its resistance to change. Finally, moving towards a normative approach, we propose a method, borrowed from bioethics, that could overcome those criticisms.

2.2 Tumor heterogeneity and its implications for clinical research

Recent technological breakthroughs, namely next-generation sequencing (NGS) and single-cell sequencing, have allowed to observe genetic variations at the single-nucleotide level between different types of tumors (*intertumor* heterogeneity), as well as between individual tumors (*intratumor* heterogeneity). Two levels of intratumor heterogeneity are generally recognized: genetic and phenotypic (Shibata and Shen 2013). For practical purposes in this chapter we are focusing on the former, leaving aside the latter. As already mentioned, with tumor heterogeneity we mean the presence of high numbers of somatic mutations within individual tumors, while some

of these are found at high allelic frequencies and are likely to appear early during tumor evolution, others are present at lower frequencies within tumor sub-clones and are acquired later (Hou et al. 2012). Tumors harboring a certain mutation could be more vulnerable and respond better to specific therapeutic agents. This variability could be exploited to develop either new combinations of drugs or even new targeted drugs, in order to provide the best therapeutic strategy for each patient. With regard to this, vemurafenib is the classic example. It has been developed in order to treat a subset of patients with late-stage melanoma harboring V600 BRAF mutation, and it has proven quite effective (Hyman et al. 2015). In the aftermath of that success, many mutations have been identified as vulnerable, druggable with new compounds or repositioning old ones. For example, in 2012 (Iyer et al. 2012) individuated two specific mutations (TSC1, NF2) in the genome of a patient with metastatic bladder cancer, who achieved a complete response after being treated with everolimus (an mTOR inhibitior). This finding showed the feasibility of exploiting tumor heterogeneity for therapeutic interest, but the patients which could benefit are few: TSC1 and NF2 genes are mutated only in a small minority of bladder cancers (Guo et al. 2013). This rarity makes testing claims of efficacy of targeted therapies particularly troublesome.

Briefly, since their introduction into drug regulation, in order to better warrant efficacy claims, phase III trials (RCTs) have become more and more rigorous, implementing robust statistical analysis, adopting hard and validated endpoints, and hardening procedures and protocols. They also became larger and larger, enrolling always more patients in order to detect even small differences and to better spot side effects (Yusuf, Collins, and Peto 1984). The downside is that nowadays RCTs are usually cumbersome experiments, expensive and costly. Nonetheless, over the years, clinical researchers and regulators have been almost blind to alternatives. Nothing has changed compared to the past: large trials (large populations for statistical power),

hypothesis testing, and control for type 1 error are still considered the yardstick. That is perfectly fine for many but not for every treatment, as for instance targeted therapies (Kirk and Hutchinson 2012; Rubin and Gilliland 2012; Sharma and Schilsky 2011; Bothwell et al. 2016).

Let us imagine a typical scenario in which some investigators want to compare a new treatment to an old one for lung cancer in a standard trial design. Suppose that the primary endpoint is the risk reduction in a negative event (e.g. death or recurrence of disease), let's say a 25% risk reduction for instance. Then suppose that the desired power of the study is 0.9, meaning that investigators want to be 90% sure of detecting that risk reduction in mortality, and with a 5% two-sided type-1 error rate (that is the standard level of statistical significance). On the basis of prior knowledge, it is legitimate to guess that the median time to the event in the control arm is 3 years. So far, no complications would arise. Now, imagine that the treatment, which the investigators would like to test, is a new compound whose efficacy has been shown in previous phases of trials only in patients who carry a specific genetic mutation, and that this population is only 3% of the total lung cancer patients. Sticking to the traditional trial design, no experiment would be feasible. Indeed, being optimistic, in that case, the accrual rate would be 4 patients per month. Then, with a minimum follow-up of 3 years, about 650 patients should be enrolled to meet the optimal sample size. This means that such a trial, in the best case, would report results in 16 years (see Berry 2015). To conduct a trial with a more manageable sample size the investigators should increase the expected risk reduction to 65% and reduce the power to a more reasonable 0.8, but such an effect is quite simply unrealistic for any cancer drug: if a drug would decrease the risk of mortality of 65% percent, it would not be even necessary to conduct an RCT to prove it. The bottom line of this thought experiment is that almost nobody would test a drug in a small population given the current regulatory

system. This is because the regulators would consider that evidence insufficient to grant market approval. As the biostatistician Donald Berry put it, "We speak of false negatives and false positives, but both are dwarfed by false neutrals, therapies that have not been and may never be evaluated in clinical trials" (Berry 2015).

2.3 Alternative trial designs

The field of trial designs research has been active since the early 80s. But just recently, the awareness of the fact that standard large RCTs are no longer an adequate tool for the evaluation of many treatments, especially new cancer drugs, has triggered the birth of many alternative approaches. In the wake of personalized medicine, much effort has been done to include the use of biomarkers in clinical research. With regard to this, for instance, a recent review of the new trial designs for therapies targeting patient subsets (Renfro et al. 2016) counted at least 7 categories of them: enrichment design; biomarker-by treatment interaction design; multi-target, multi-agent biomarker strategy design; biomarker design with response-adaptive randomization; umbrella trial; and basket trial. However, they all share one common feature: they employ an "adaptive design". This latter is a label for those approaches that allow for adaptations of the study protocol while the trial is still ongoing, on the basis of the information collected and generated during the trial. "Adaptations" can be various and include, for instance, changes in groups size, adjustment in treatment dosage, or even dropping an arm of the study. These features make this kind of trials appropriate for testing cancer drugs since they can also include a biomarker to identify the class of patient that would benefit. Adaptive designs deliver many other advantages. First of all, adaptive approaches can reduce the sample size needed to detect the supposed effect, therefore reducing also the cost and time needed to complete the trial. Second, adaptive designs are also considered more ethical than standard RCTs, since they minimize the chance

of exposure to a therapy that is not effective[2]. Lastly, adaptive designs have a pragmatic sharp-edge: they can shorten the drug development and the approval path, and this can incentive drug companies to invest more in the drug discovery endeavor.

However, some serious critiques to adaptive trials may be listed. The most pressing one is that, so far, every single "adaptive trial" has employed a different study protocol. They can also require indifferently either a frequentist or a Bayesian statistical analysis, these latter being particularly complex. So, since these study protocols can be very intricate and cumbersome, "clear and pervasive evidence would recede behind the [statistical] veil" (Caplan, Plunkett, and Levin 2015). Moreover, given their adaptive nature, such experiments can hardly be replicable: of course, the information generated and collected in a trial can be completely different from that generated in another trial, despite the adoption of the same initial protocol. This non-replicability makes them suspect. Indeed, the FDA guidelines on adaptive trials clearly eschew those designs that are not “well-understood” (*ibid*).

Usually, defenders of RCTs argue that any trial which does not include a statistical significance test will never produce “objective” knowledge. This belief is based on the shared perception that the classical/frequentist test of statistical significance is an objective and impartial assessment of a trial’s result. Although, this granted epistemic superiority is quite controversial (Teira 2011). But the most pressing reason explaining regulators’ reluctance towards endorse adaptive trial designs is the potential lack of a yardstick. FDA experts, for instance, are aware of the limitations of RCTs and frequentist statistical inference. However, they are willing to pay this price in order to avoid a pluralism of criteria and methods that would be

[2] The ethical preeminence of adaptive trials is questionable. See for instance (Hey and Kimmelman 2015).

practically impossible to manage. As Robert Temple (Deputy Director for Clinical Sciences at the FDA) once put it:

> *"Of course, everybody knows that "P < 0.05"] is sort of stupid. Why should it always be the same? Why shouldn't it be adjusted to the situation, to the risks of being wrong in each direction? The alternative to adopting a standard is to actually determine a criterion for success on the spot for each new case. That is my idea of a nightmare. So, we use a foolish, if you like, simplification. […] I don't want to have a symposium for every new trial to decide on an acceptable level of evidence. My point is that all of these things need to be well enough understood so we can actually implement procedures that won't drive everybody crazy, that won't involve constant arguments about the strength or nature of each assumption every time" (Berry et al. 2005).*

Temple does not neglect the epistemic value of methodological pluralism, and he acknowledges that having just one way to evaluate and approve a new drug is just a sub-optimal procedure. Yet, the aim of FDA is not to increase scientific knowledge, giving a reliable answer to a scientific question, but rather guarantee to consumers both access to drugs and protection from severe adverse effects, with the final aim of improving health in the general population. Then, paradoxically, lack of standardization could slow down the path for approval rather than speed it up, and could open the door to intentional bias.

Generally speaking, in the field of methodology of clinical research two views emerged: that of "conservatives" (Cox, Borio, and Temple 2014; Joffe 2014; Rid and Emanuel 2014; Nelson et al. 2015) who uphold the use of the gold-standard (RCTs) in all clinical research without exception; and the view of "reformists" (Caplan, Plunkett, and Levin 2015; Adebamowo et al. 2014; Simon et al. 2015; Hohl 2015), who instead

consider the use of alternative trial designs justified under some circumstances. There are both epistemological and ethical reasons supporting the two positions.

At the basis of the conservatives' argument lay some well-known reasons, endorsed especially by the mainstream fringe of the medical community: RCTs are the only experimental design that justifies an objective and reliable scientific inference. This is because in a double-blind RCT we can put in place a series of controls in order to prevent any bias that could spoil the validity of the result. In particular, randomization with a control group is held to be the best way to minimize confounding factors. Moreover, since the use of human subjects in clinical research is justified only to the extent to which the experiment provides a reliable answer to the scientific question under investigation, so that the result can provide a clear benefit for all the other people, RCTs have also a legitimization from an ethical point of view. For instance, conservatives claim that without a concurrent randomized control group, the evidence concerning drug efficacy and safety will be compromised and the trial might produce misleading results. That is also why other trial designs are not only methodologically flawed but also unethical. In few words, alternative designs might do more harm than good to future patients.

Resistance to change by the conservatives is due especially to a "belief that in any research endeavor one must always test a null hypothesis with adequate control of the type 1 error rate concerning declaring false positive results" (Caplan, Plunkett, and Levin 2015). This explains also the general reluctance to adopt Bayesian statistics, for instance. We do not want to engage the never-ending debate between Bayesians and frequentists about the foundations of statistical inference, but we want to highlight a simple and straightforward fact, that there are many cases of "inquiries" in which testing a null hypothesis simply makes no sense at all. Consider, for instance, a coach who must select the best football players for his team. It does not make any sense to

test the null-hypothesis that "no player is any better than any other player – or worse, testing the null hypothesis that Johnny can't [kick] better than a placebo" (*ibid*.). Without going so far away, even when one must decide to choose which investment put in his own portfolio testing the null sounds pretty ridiculous. This is because the goals are completely different, what we need in cases like these is just “to select the best option and move forward” (*ibid*.). The goodness of the choice does not rely on the method by which that choice has been made, rather it is assessed by means of other criteria, saying, for instance, empirical success. As we already suggest elsewhere (see chapter 3) drug withdrawals for safety reasons are an appropriate empirical benchmark to assess the quality of regulatory decision making. Instead, coming back to our fictional examples, the coach who will win the league has likely made the best choices.

From the 60s on, in biomedical research, RCTs became the gold standard methodology for fulfilling many different tasks, from taking decisions regarding the licensing of new drugs, to informing medical practice (e.g. EBM). In all these cases RCTs have shown to meet some practical challenges, which render their results quite useless for the aims they are intended for. We have already seen some of their practical limitations for testing a special class of cancer drugs, but it is almost the same for preventive research (Golfam et al. 2015), or for neurosurgery interventions (Mansouri et al. 2015). Time and again, new experimental designs have been proposed in order to overcome the limits of the RCTs in order to get a meaningful answer to the question under investigation. The idea is quite convincing: disregarding statistical considerations, following stubbornly a recipe for which one cannot get the right ingredients is irrational, to say the least. For instance, as many epidemiologists and philosophers of science have often argued (Vandenbroucke 2004, 2008; Osimani 2013) (Vandenbroucke 2004, 2008, Osimani 2013, Russo 2014), under certain circumstances

a well-designed observational study can provide a better answer to a scientific question rather than a poorly designed and performed randomized controlled trial. As Federica Russo (Russo 2014) put it, picking up an old British adage, “the proof of the pudding is in the eating, not in the gold standard recipe”. This is to say that the quality of regulatory decision making does not depend only on the methods which have been employed in the process.

2.4 A casuistic approach to cancer clinical trials

Recently, philosophers of science (Hey and Kesselheim 2016) have argued that the gold standard for testing treatments presupposes agnosticism about underlying biological theories. Indeed, as very often it has been the case, RCTs can effectively test the efficacy of some drugs, even without a theoretical understanding of the biological mechanism explaining why the drug is effective. As described in a masterly manner by (Keating and Cambrosio 2012), since the beginning of cancer pharmacological research, researchers and research institutions, such as NIH, have employed a high-throughput approach: screening for those chemicals and natural compounds among many, which could have a cytotoxic effect *in vitro*. That was basically the mainstream line of research until the 90s, when genetics and genomics had a great boost due to many important scientific breakthroughs, the achievement of the Human Genome Project in 2003 above all. As a result, in the era of personalized medicine, there is a clear biological hypothesis under testing in a clinical trial, for instance: “Treatment T is effective for condition C, as defined by testing positive for biomarker B, where B is determined by diagnostic assay A” (Hey and Kesselheim 2016). But this hypothesis encompasses different additional assumptions, such as "why A is a reliable test for B" or "why B should predict activity of T against C" (*ibid*.). These additional assumptions are tested long before the hypothesis reaches the phase of clinical testing. If the assumptions hold, then it might

be reasonable to even accept less robust statistical evidence, since it is supported by an underlying biological understanding. From this perspective, it is hard to argue that if we dispense with RCTs then we are decreasing the evidential standard and putting future patients at more risk. However, only a proper appraisal of the strength of evidence (biological and statistical) in each single case can allow reliable regulatory decision making. To this regard, it seems inevitable for the FDA to discuss and deliberate on every single new trial. Yet, this does not mean that the approval process would necessarily turn into a "nightmare". It might be just a matter of changing the approach to drug regulation for a special class of drugs and invest on it. There is no need to have a symposium for every registration trial in order to reach a consensus and a decision, it might be sufficient to adopt a *casuistic* approach.

Let me briefly illustrate it. The casuistry is a very old approach to moral decision-making and practical ethics. There are some traces of it already in the Talmud, particularly in the hermeneutical rules to interpret the precepts of the Pentateuch. Some elements of the casuistry method are present also in the work of Cicero (*De Officis, De inventione*), but mostly in Aristotle's *Rhetoric*, which inspired greatly the casuistry of the late medieval and early modern age. In this latter period, the casuistry took on a more organized shape thanks to the work of the Jesuits. The method was "invented" in order to solve a pressing issue at that time: the problem of "doubtful conscience", i.e. what to do in all those cases in which it is not clear what is the best moral decision to make. People of medieval and early modern age were used to ask for advice to their confessors, who were often in trouble in giving the right advice. In that context, the casuistry method consisted of the technique (or art) of giving answers to specific moral questions without justifying them through moral principles, but just using an approach which we could define "topical" or "comparative-inductive". The idea was to assume an index case, which was sharp and clear, and then to analyze the

new cases by analogical reasoning. The big advantage of casuistry was that it did not require any agreement on ethical theories or moral principles, but just a consensus on the conclusions of the index cases. This method has been on the edge until the second half of the seventh century, when it was harshly criticized especially by Jansenist philosophers, such as for instance Blaise Pascal.

More recently, in the context of contemporary debate on methods of bioethics, the casuistry has been recovered by Jonsen and Toulmin (Jonsen and Toulmin 1988). The starting point is the criticism to the Beauchamp and Childress' leading bioethical paradigm (Beauchamp and Childress 2001). According to Jonsen and Toulmin the method of the principles, which deductively infer moral conclusions from accepted general moral principles, is too rigid and dogmatic. The two bioethicists then suggest recovering the casuistry method, since it implies less demanding assumptions. In fact, as just mentioned, it is not a top-down approach, but it requires accepting a conclusion of an index case and then comparing a new case to it, according to similarities and differences, in order to assess whether the new case can be part of the same class of the index one or not. The idea is that it is much easier to agree on an index case rather than accepting general moral principles. Following this sort of reasoning, it is possible to reach a conclusion which is only "probable" and not "certain". Nonetheless, sometimes it is more convenient to set the bar a little bit lower, in order to pursue the aim of reaching a large consensus on a moral case. In the index cases the conclusions are usually very sharp and indisputable. For instance, there is no doubt that one is morally obligated to return money to a friend who lent him some. However, is not equally clear that one must return to a friend a weapon with which this latter wants to commit murder. In this latter case, we assist a common conflict between moral rules/norms: returning money lent, and not being an accessory to murder. The strength of the casuistry method is precisely being *theory-independent* (i.e. moral

judgments are not validly deduced from the universal principles of this and that ethical theory), hence more suitable to generate a wide consensus. Jonsen and Toulmin got inspiration also from the experience of the National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research that in the 70s, in the aftermath of the Tuskegee scandal, was called upon to establish the ethical limits and set the policies of biomedical research in the U.S (Varmus and Satcher 1997). The members of the Commission had varied backgrounds and expertise (they included philosophers, psychologists, theologians, medical scientists, etc.) representing different categories of stakeholders (men and women, Catholics, Protestants, Jews, etc.). Because of those differences, before the Commission's work started, few observers expected the members to find an agreement either on general principles or on particular problems. Nevertheless, so long as the commissioners stood on the cases, they generally agreed on their practical conclusions. Instead, when they explained the reasons justifying their judgments, an open disagreement emerged. The philosophical wonder stems from the fact that in the Commission's discussions the locus of moral certitude did not lie in general principles but rather "in a shared perception of what was specifically at stake in particular kinds of human situations" (Jonsen and Toulmin 1988). It is clear that this perspective moral decision making requires a sort of "practical wisdom" (*phronesis*, a concept notoriously borrowed from Aristotle's Nicomachean Ethics): the more this capacity, the better the judgments.

Broadly speaking, over the years, bioethicists have largely preferred the rival method of the principles, as many have raised serious questions about the justification power of casuistry, despite its appeal. Often, casuists rely on social consensus to decide the paradigmatic cases, and critics argue that such a consensus is hardly reachable in the real-world scenario. Then, if there is not sufficient warranty for the decision about the paradigmatic cases, casuistic arguments are epistemically defective. Moreover,

even if we could agree about the paradigmatic case, one can still deny that the case under investigation is similar in any meaningful way, bringing out the dissimilarities between the two cases. So, lack of consensus would undermine also the casuistic based moral reasoning. Considering this, we think it is fair to say that the casuistic reasoning is epistemically vulnerable in many cases. Although this does not render casuistry useless, "it shows that its conclusions are tentative at best" (Spielthenner 2016). A consensus on moral decisions is hardly reachable despite the method because moral cases are often very blurry. Moreover, one could argue that in the interpretation of moral cases one cannot get along without an ethical theory.

Nonetheless, in the regulatory decision setting, we are dealing with a rather different kind of cases, on which we can fairly easily suspend our judgment on principles. For instance, frequentist and Bayesian statisticians would hardly agree about the foundations of statistical inference and the epistemic import of different study designs. Indeed, the relative merit of the Bayesian and frequentist approaches continues to be the subject of debate. However, on the one hand, Bayesians do not hold that all the regulatory decisions based on results from standard RCTs are poor, just because they require a frequentist framework. Otherwise, they should be skeptical also about almost all the results of every scientific discipline, since frequentism is the current dominant paradigm. On the other hand, frequentist statisticians do not deny the reliability of many studies that employ a Bayesian statistical analysis. For example, in 2003 the Center for Drugs and Experimental Research of the FDA approved Pravigard Pac (Bristol-Myers Squibb) based on Bayesian analyses of efficacy (Berry 2006; Jack Lee and Chu 2012), and so far no one has complained about that decision, meaning that, once it reached the market, the drug efficacy did not disappear nor showed unexpected severe adverse effects. Therefore, we think that reaching a consensus on some paradigmatic cases for regulatory decisions is not only possible but

also desirable. Ideally, this would allow the FDA experts to exploit the methodological pluralism, shorten the approval path for drug clinical development, and satisfy the interests of all the stakeholders. The casuistic approach would be an intermediate way between the current one-size-fits-all approach, which has been shown to be inadequate for testing molecularly targeted drugs, resolving the risk of having a symposium for every new trial to decide on an acceptable level of evidence, which would sink the FDA into a decisional quagmire.

2.5 Potential paradigmatic case: crizotinib

As stated above, the conservatives would never agree with the reformists about the possibility of getting a reliable result from a study design that does not provide a control group. There is no way by which they can agree on that on the basis of their theories about experimental design and statistical inference. Let us consider then the story of clinical development and approval of crizotinib as a potential paradigmatic case that we think might reconcile the conservatives and reformists.

Crizotinib is a small-molecule that acts competitively inhibiting the ATP pocket of ALK, MET, and ROS1 kinases (Selaru et al. 2016). In physiological conditions, kinases provide proliferation signals phosphorylating target proteins. When mutated, they trans-autophosphorylate and result in potent tumorigenic drivers that activate downstream oncogenic signals, leading to aberrant cell proliferation and survival. The inhibition of those genes should then block this mechanism of action. Crizotinib was originally synthesized in 2005 precisely as an inhibitor of one of those kinases: the hepatocyte growth factor receptor (HGFR). Subsequently, it was found to inhibit phosphorylation of NPM-ALK in anaplastic large-cell lymphoma cells (ALCL). Also, ALK kinases have been found chromosomally rearranged in non-small cell lung cancer (NSCLC) with EML4 (Soda et al. 2007)

Crizotinib received accelerated approval by the FDA in August 2011 for the treatment of patients with locally advanced or metastatic ALK-positive NSCLC (Kazandjian et al. 2014), basically 6 years after the initial discovery of the molecule, that is a very short time considering that in general a drug takes about 10-15 years to enter the market. The crizotinib approval was based on the data from two open-label, single arm, phase I and II clinical trials: PROFILE 1001 and PROFILE 1005 (Ou 2011), respectively. Initially, the former, as all the phase I trials, recruited patients with various advanced solid tumors resistant to standard chemotherapy, in order to determine the maximum tolerated dose (MTD) of the drug. Once established the MTD, the study evaluated the safety and antitumor activity, first in patients who harbored MET amplifications (Ou et al. 2012) Then, after the discovery of ALK gene rearrangements in 2007 (Soda et al. 2007), a cohort of patients with ALK-positive NSCLC was added to the study in 2008. Since it was a phase I trial, PFS (progression free survival) was not one of the main outcomes, although the median PFS of 10 months was astonishing considering that in general, with standard therapy, it is about 3 months. Moreover, crizotinib was well tolerated with minor adverse events. So, on the basis of these early promising results, in 2010 a phase II study started. The PROFILE 1005 primary objectives were antitumor activity and safety, but, as all the phase II trials, the investigators began to collect also some efficacy outcomes in addition to PFS. The 136 patients enrolled had an ORR (objective response rate)[3] of 51% and a duration response rate of 41.9 weeks (Crino et al. 2011). Furthermore, safety data were consistent with those observed in PROFILE 1001. As stated above, in 2011 crizotinib obtained the FDA accelerated approval based on these data.

[3] The FDA defines ORR as the proportion of patients with a tumor size reduction of a predefined amount and for a minimum period of time.

Clinical development of crizotinib had to face many challenges, not least the small population of the ALK-positive NSCLC, which is, as explained above, one of the major obstacles in testing targeted (or personalized) treatments. Moreover, historical data on typical end points in that specific population as well as information on natural history of the diseases were lacking: because the relevance of ALK mutation was unknown, physicians never screened patients for that mutation. This made the interpretation of the results of the crizotinib trials particularly challenging since there was not any control arm nor any historical controls to compare with. In order to overcome this problem, investigators performed some simulations to predict the outcomes of virtual controls (see Selaru et al. 2016).

We firmly believe that even the most reactionary defender of RCTs could not neglect the epistemic import of the crizotinib trials nor could deny the validity of their results. So, in principle, they could also agree that under specific conditions a small trial - without any control group - can effectively lead to a right regulatory decision. These conditions may include:

- knowledge of the mechanisms of action of the drug supported by a strong biological rationale, plus the description of the patient population;
- the observation of a rapid response to therapy and durable anti-tumor activity, that is consistent with the biological rationale;
- a favorable safety profile, meaning the observation of grade 1-2 adverse events only.

Our intuition is that it is far easier to reach an agreement on these conditions than, for instance, on the statistical validity of the covariate-matched (Rubin 2006) analysis and covariate-adjusted analysis (Tian et al. 2012) employed by the investigators in order to overcome the absence of controls. When it comes to

regulatory decisions, experts should understand the difference between “theory” and “practice”, that is, “between the demands of scientific understanding and those of practical good” (Jonsen e Toulmin 1988).

2.6 Conclusion

In the light of the methodological revolution driven by the discovery of tumor heterogeneity, we have contended the more general idea that scientific inference and regulatory decision-making can be reduced to sound statistical inferences and to go-no-go choices, respectively. In detail, we have shown how the current gold standard for drug testing is inadequate for new oncological targeted (or personalized) treatments, mostly due to practical challenges such as the difficulty to enroll enough patients to reach a sufficient statistical power. For this reason, novel experimental designs are under investigation, adaptive trials being the most promising. However, regulatory agencies have not fully endorsed them yet because they are afraid that a lack of standardization could actually transform the decision-making approval into a never-ending process. Evaluating applications for drug approval case by case can further slow down the process of drug development, which instead must be accelerated. With this regard, we proposed a third way, borrowed from bioethics, to avoid such an impasse while dispensing with a strict regulatory standard: the casuistry. In the casuistic framework, which conceptually embodies methodological pluralism, experts must agree not on theories behind the features of every single trial design, but only on single real cases. Once they found an agreement on a small set of paradigmatic cases, they can easily assess new ones by analogical reasoning. Finally, we proposed as a potential paradigmatic case the clinical development of crizotinib, a molecularly targeted drug for the treatment of lung cancer recently approved on the basis of two small single-arm early phase trials.

If post-genomics will fully deliver its promises, the future of clinical trials will necessarily be much more different. The legislator then must be prepared to rise to the regulatory challenge.

References

Adebamowo, Clement, Oumou Bah-Sow, Fred Binka, Roberto Bruzzone, Arthur Caplan, Jean-François Delfraissy, David Heymann, et al. 2014. "Randomised Controlled Trials for Ebola: Practical and Ethical Issues." *The Lancet* 384 (9952):1423–24. https://doi.org/10.1016/S0140-6736(14)61734-7.

Beauchamp, Tom L., and James F. Childress. 2001. *Principles of Biomedical Ethics*. Oxford University Press.

Berry, Donald A. 2006. "Bayesian Clinical Trials." *Nature Reviews Drug Discovery* 5 (1):nrd1927. https://doi.org/10.1038/nrd1927.

———. 2015. "The Brave New World of Clinical Cancer Research: Adaptive Biomarker-Driven Trials Integrating Clinical Practice with Clinical Research." *Molecular Oncology* 9 (5):951–59. https://doi.org/10.1016/j.molonc.2015.02.011.

Berry, Donald A, Steven N Goodman, and Thomas A Louis. 2005. "Floor Discussion." *Clinical Trials* 2 (4):301–4. https://doi.org/10.1191/1740774505cn101oa.

Bertolaso, Marta. 2011. "Hierarchies and Causal Relationships in Interpretative Models of the Neoplastic Process." *History and Philosophy of the Life Sciences* 33 (4):515–35.

Boniolo, Giovanni, and Marco J. Nathan. 2016. *Philosophy of Molecular Medicine: Foundational Issues in Research and Practice*. Taylor & Francis.

Bothwell, Laura E., Jeremy A. Greene, Scott H. Podolsky, and David S. Jones. 2016. "Assessing the Gold Standard — Lessons from the History of RCTs." *New England Journal of Medicine* 374 (22):2175–81. https://doi.org/10.1056/NEJMms1604593.

Caplan, Arthur L., Carolyn Plunkett, and Bruce Levin. 2015. "Selecting the Right Tool For the Job." *The American Journal of Bioethics* 15 (4):4–10. https://doi.org/10.1080/15265161.2015.1010993.

Cox, Edward, Luciana Borio, and Robert Temple. 2014. "Evaluating Ebola Therapies — The Case for RCTs." *New England Journal of Medicine* 371 (25):2350–51. https://doi.org/10.1056/NEJMp1414145.

Crinò, L., D. Kim, G. J. Riely, P. A. Janne, F. H. Blackhall, D. R. Camidge, V. Hirsh, et al. 2011. "Initial Phase II Results with Crizotinib in Advanced ALK-Positive Non-Small Cell Lung Cancer (NSCLC): PROFILE 1005." *Journal of Clinical Oncology* 29 (15_suppl):7514–7514. https://doi.org/10.1200/jco.2011.29.15_suppl.7514.

Diaz Jr, Luis A., Richard T. Williams, Jian Wu, Isaac Kinde, J. Randolph Hecht, Jordan Berlin, Benjamin Allen, et al. 2012. "The Molecular Evolution of Acquired Resistance to Targeted EGFR Blockade in Colorectal Cancers." *Nature* 486 (7404):537–40. https://doi.org/10.1038/nature11219.

Germain, Pierre-Luc. 2012. "Cancer Cells and Adaptive Explanations." *Biology & Philosophy* 27 (6):785–810. https://doi.org/10.1007/s10539-012-9334-2.

Golfam, Mohammad, Reed Beall, Jamie Brehaut, Sara Saeed, Clare Relton, Fredrick D. Ashbury, and Julian Little. 2015. "Comparing Alternative Design Options for Chronic Disease Prevention Interventions." *European Journal of Clinical Investigation* 45 (1):87–99. https://doi.org/10.1111/eci.12371.

Guo, Yanan, Yvonne Chekaluk, Jianming Zhang, Jinyan Du, Nathanael S Gray, Chin-Lee Wu, and David J Kwiatkowski. 2013. "*TSC1* Involvement in Bladder Cancer: Diverse Effects and Therapeutic Implications: *TSC1* in Bladder Cancer." *The Journal of Pathology* 230 (1):17–27. https://doi.org/10.1002/path.4176.

Hey, Spencer Phillips, and Aaron S. Kesselheim. 2016. "Countering Imprecision in Precision Medicine." *Science* 353 (6298):448–49. https://doi.org/10.1126/science.aaf5101.

Hey, Spencer Phillips, and Jonathan Kimmelman. 2015. “Are Outcome-Adaptive Allocation Trials Ethical?” *Clinical Trials* 12 (2):102–6. https://doi.org/10.1177/1740774514563583.

Hohl, Rj. 2015. “Oncology Trial Design: More Accurately and Efficiently Advancing the Field.” *Clinical Pharmacology & Therapeutics* 97 (5):430–32. https://doi.org/10.1002/cpt.94.

Hou, Yong, Luting Song, Ping Zhu, Bo Zhang, Ye Tao, Xun Xu, Fuqiang Li, et al. 2012. “Single-Cell Exome Sequencing and Monoclonal Evolution of a JAK2-Negative Myeloproliferative Neoplasm.” *Cell* 148 (5):873–85. https://doi.org/10.1016/j.cell.2012.02.028.

Hyman, David M., Igor Puzanov, Vivek Subbiah, Jason E. Faris, Ian Chau, Jean-Yves Blay, Jürgen Wolf, et al. 2015. “Vemurafenib in Multiple Nonmelanoma Cancers with BRAF V600 Mutations.” *New England Journal of Medicine* 373 (8):726–36. https://doi.org/10.1056/NEJMoa1502309.

Iyer, Gopa, Aphrothiti J. Hanrahan, Matthew I. Milowsky, Hikmat Al-Ahmadie, Sasinya N. Scott, Manickam Janakiraman, Mono Pirun, et al. 2012. “Genome Sequencing Identifies a Basis for Everolimus Sensitivity.” *Science* 338 (6104):221–221. https://doi.org/10.1126/science.1226344.

Jack Lee, J., and Caleb T. Chu. 2012. “Bayesian Clinical Trials in Action.” *Statistics in Medicine* 31 (25):2955–72. https://doi.org/10.1002/sim.5404.

Joffe, Steven. 2014. “Evaluating Novel Therapies During the Ebola Epidemic.” *JAMA* 312 (13):1299–1300. https://doi.org/10.1001/jama.2014.12867.

Jonsen, Albert R., Stephen Edelston Toulmin, and Stephen Toulmin. 1988. *The Abuse of Casuistry: A History of Moral Reasoning*. University of California Press.

Kazandjian, Dickran, Gideon M. Blumenthal, Huan-Yu Chen, Kun He, Mona Patel, Robert Justice, Patricia Keegan, and Richard Pazdur. 2014. “FDA Approval Summary: Crizotinib for the Treatment of Metastatic Non-Small Cell Lung Cancer With Anaplastic Lymphoma Kinase Rearrangements.” *The Oncologist* 19 (10):e5–11. https://doi.org/10.1634/theoncologist.2014-0241.

Keating, Peter, and Alberto Cambrosio. 2011. *Cancer on Trial: Oncology as a New Style of Practice*. University of Chicago Press.

Kirk, Rebecca, and Lisa Hutchinson. 2012. "Oncology Trials—the Elephant in the Room." *Nature Reviews Clinical Oncology* 9 (4):nrclinonc.2012.33. https://doi.org/10.1038/nrclinonc.2012.33.

Mansouri, Alireza, Samuel Shin, Benjamin Cooper, Archita Srivastava, Mohit Bhandari, and Douglas Kondziolka. 2015. "Randomized Controlled Trials and Neuro-Oncology: Should Alternative Designs Be Considered?" *Journal of Neuro-Oncology* 124 (3):345–56. https://doi.org/10.1007/s11060-015-1870-6.

Nelson, Robert M., Michelle Roth-Cline, Kevin Prohaska, Edward Cox, Luciana Borio, and Robert Temple. 2015. "Right Job, Wrong Tool: A Commentary on Designing Clinical Trials for Ebola Virus Disease." *The American Journal of Bioethics* 15 (4):33–36. https://doi.org/10.1080/15265161.2015.1010018.

Osimani, Barbara. 2013. "The Precautionary Principle in the Pharmaceutical Domain: A Philosophical Enquiry into Probabilistic Reasoning and Risk Aversion." *Health, Risk & Society* 15 (2):123–43. https://doi.org/10.1080/13698575.2013.771736.

Ou, Sai-Hong Ignatius. 2011. "Crizotinib: A Novel and First-in-Class Multitargeted Tyrosine Kinase Inhibitor for the Treatment of Anaplastic Lymphoma Kinase Rearranged Nonsmall Cell Lung Cancer and Beyond." *Drug Design, Development and Therapy*, November, 471. https://doi.org/10.2147/DDDT.S19045.

Ou, Sai-Hong Ignatius, Cynthia Huang Bartlett, Mari Mino-Kenudson, Jean Cui, and A. John Iafrate. 2012. "Crizotinib for the Treatment of *ALK* -Rearranged Non-Small Cell Lung Cancer: A Success Story to Usher in the Second Decade of Molecular Targeted Therapy in Oncology." *The Oncologist* 17 (11):1351–75. https://doi.org/10.1634/theoncologist.2012-0311.

Pirmohamed, Munir. 2001. "Pharmacogenetics and Pharmacogenomics." *British Journal of Clinical Pharmacology* 52 (4):345–47. https://doi.org/10.1046/j.0306-5251.2001.01498.x.

Plutynski, Anya. 2013. “Cancer and the Goals of Integration.” *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences* 44 (4, Part A):466–76. https://doi.org/10.1016/j.shpsc.2013.03.019.

Renfro, Lindsay A., Himel Mallick, Ming-Wen An, Daniel J. Sargent, and Sumithra J. Mandrekar. 2016. “Clinical Trial Designs Incorporating Predictive Biomarkers.” *Cancer Treatment Reviews* 43 (Supplement C):74–82. https://doi.org/10.1016/j.ctrv.2015.12.008.

Rid, Annette, and Ezekiel J. Emanuel. 2014. “Ethical Considerations of Experimental Interventions in the Ebola Outbreak.” *The Lancet* 384 (9957):1896–99. https://doi.org/10.1016/S0140-6736(14)61315-5.

Rubin, Donald B. 2006. *Matched Sampling for Causal Effects*. Cambridge University Press.

Rubin, Eric H., and D. Gary Gilliland. 2012. “Drug Development and Clinical Trials—the Path to an Approved Cancer Drug.” *Nature Reviews Clinical Oncology* 9 (4):nrclinonc.2012.22. https://doi.org/10.1038/nrclinonc.2012.22.

Russo, Federica. 2014. “What Invariance Is and How to Test for It.” *International Studies in the Philosophy of Science* 28 (2):157–83. https://doi.org/10.1080/02698595.2014.932528.

Selaru, P, Y Tang, B Huang, A Polli, Kd Wilner, E Donnelly, and Dp Cohen. 2016. “Sufficiency of Single-Arm Studies to Support Registration of Targeted Agents in Molecularly Selected Patients with Cancer: Lessons from the Clinical Development of Crizotinib: Clinical Development of Crizotinib.” *Clinical and Translational Science* 9 (2):63–73. https://doi.org/10.1111/cts.12388.

Sharma, Manish R., and Richard L. Schilsky. 2011. “Role of Randomized Phase III Trials in an Era of Effective Targeted Therapies.” *Nature Reviews Clinical Oncology* 9 (4):nrclinonc.2011.190. https://doi.org/10.1038/nrclinonc.2011.190.

Shibata, Maho, and Michael M. Shen. 2013. “The Roots of Cancer: Stem Cells and the Basis for Tumor Heterogeneity.” *BioEssays* 35 (3):253–60. https://doi.org/10.1002/bies.201200101.

Simon, R, Gm Blumenthal, Ml Rothenberg, J Sommer, Sa Roberts, Dk Armstrong, Lm LaVange, and R Pazdur. 2015. “The Role of Nonrandomized Trials in the Evaluation of Oncology Drugs.” *Clinical Pharmacology & Therapeutics* 97 (5):502–7. https://doi.org/10.1002/cpt.86.

Simon, Richard. 2010. “Translational Research in Oncology: Key Bottlenecks and New Paradigms.” *Expert Reviews in Molecular Medicine* 12 (October). https://doi.org/10.1017/S1462399410001638.

Soda, Manabu, Young Lim Choi, Munehiro Enomoto, Shuji Takada, Yoshihiro Yamashita, Shunpei Ishikawa, Shin-ichiro Fujiwara, et al. 2007. “Identification of the Transforming EML4–ALK Fusion Gene in Non-Small-Cell Lung Cancer.” *Nature* 448 (7153):561–66. https://doi.org/10.1038/nature05945.

Spielthenner, Georg. 2016. “The Casuistic Method of Practical Ethics.” *Theoretical Medicine and Bioethics* 37 (5):417–31. https://doi.org/10.1007/s11017-016-9382-8.

Teira, David. 2011. “Bayesian Versus Frequentist Clinical Trials.” In *Philosophy of Medicine [Handbook of Philosophy of Science, Vol. 16],* edited by Gifford Fred. Elsevier.

Tian, L., T. Cai, L. Zhao, and L.-J. Wei. 2012. “On the Covariate-Adjusted Estimation for an Overall Treatment Difference with Data from a Randomized Comparative Clinical Trial.” *Biostatistics* 13 (2):256–73. https://doi.org/10.1093/biostatistics/kxr050.

Vandenbroucke, Jan P. 2004. “When Are Observational Studies as Credible as Randomised Trials?” *The Lancet* 363 (9422):1728–31. https://doi.org/10.1016/S0140-6736(04)16261-2.

Vandenbroucke, Jan P. 2008. “Observational Research, Randomised Trials, and Two Views of Medical Science.” *PLOS Medicine* 5 (3):e67. https://doi.org/10.1371/journal.pmed.0050067.

Varmus, Harold, and David Satcher. 1997. “Ethical Complexities of Conducting Research in Developing Countries.” *New England Journal of Medicine* 337 (14):1003–5. https://doi.org/10.1056/NEJM199710023371411.

Yusuf, Salim, Rory Collins, and Richard Peto. 1984. “Why Do We Need Some Large, Simple Randomized Trials?” *Statistics in Medicine* 3 (4):409–20. https://doi.org/10.1002/sim.4780030421.

3. RULES VERSUS STANDARDS: A LEGAL-PHILOSOPHICAL FRAMEWORK FOR DRUG REGULATION

"It is as if we had hardened the empirical proposition into a rule. And now we have, not an hypothesis that gets tested by experience, but a paradigm with which experience is compared and judged. And so a new kind of judgment" (Ludwig Wittgenstein, Remarks on the Foundations of Mathematics, VI:22)

3.1 Introduction

During this last decade, philosophers of science have addressed the use of scientific evidence for policy-making purposes in many different ways. Such discussions have often adopted a Platonic stance: epistemology goes first. Given a policy-making problem, the philosopher should just identify which sort of evidence will better solve it. But the policy-making process may often be more complex. Our goal in this chapter is to show the limitations of this Platonic approach, advocating instead for a more experimental take: rather than just criticizing the evidence actually used on purely a priori grounds and present a principled alternative, philosophers should also care about the empirical benchmarks that would allow the public to see whether their principled alternatives work better than the already established approach.

Our case in point is going to be drug regulation, in which for the last five decades RCTs have been used to test the safety and efficacy of new compounds. Philosophers of science have extensively criticized the epistemic superiority of clinical trials for such

purpose and have defended various forms of evidential pluralism, on a priori epistemic grounds. We are going first to restate this epistemic debate in legal terms (section 2): for regulatory purposes, evidential pluralism implies a *standard*-based decision-making process, whereas the current approach to drug testing is based instead on *rules*. As of today, most regulatory decisions operate on a simple rule: if there are two positive RCTs, approve the drug. Philosophers of science contend instead that regulators would do better if they addressed the substantive question of whether a drug is safe and effective with the best evidence available. This is a standard-based approach.

Legal philosophers have extensively analyzed the advantages of rules and standards in terms of their costs. We will discuss the cognitive and practical costs involved in making a regulatory decision about drug approvals, showing how much more costly standards are. We will then argue that good regulatory decisions are not just a matter of the epistemic approach implemented, but depend on the actual circumstances in which such decisions are taken. Whereas rules can be implemented without substantive discussion, standards require deliberating committees. A number of regulatory decisions at the FDA are actually taken by committees deliberating on a standard-basis. It has been shown that they are, at least, as vulnerable to conflicts of interest as RCT-based rules, if not more. In addition to the costs of using rules or standards, we should take into consideration the further costs of protecting the regulatory decision-making process from third-party influences.

With all this in sight, we proceed to propose an empirical benchmark to test the superiority of any standard-based approach to the RCT-rule: the number of drugs withdrawn from the market for safety reasons under the current institutional setup of the FDA. We discuss the estimates available for setting this benchmark and its sensitivity to regulatory capture. RCT-based decisions at the FDA have reached a low

error rate with a reasonable degree of impartiality. In presenting an alternative, philosophers should discuss not just the possibility of improving upon this benchmark, but also the costs of achieving such an improvement. The public should decide whether the improvement is worth those costs.

3.2 Rules versus standards in drug regulation

Let us first introduce the distinction between rules and standards, following an extensive literature in legal and political studies. Both are *legal directives* with a conditional structure of the form "If X, then Y". The antecedent is the "trigger": an event plus some concomitant circumstances. The consequent is the "response". Rules have a hard empirical trigger, and a hard determinate response. Standards have a soft evaluative trigger, and a soft guided response (Schlag 1985).

"Dogs are not allowed in bars" is a rule of the form "If x is a dog, x is not allowed in a bar". The trigger is an unambiguous fact (either x is a dog or not); the response is not open to interpretation. Rules have justifications: the reason why they were originally established. Preventing dogs from entering bars might have been justified in order to avoid disturbances to customers. According to (Schauer 1993), rules should be always implemented independently of their justification: seeing-eye dogs are surely trained not to disturb anyone in public spaces, but they are dogs nonetheless and, according to the rule, they should not be allowed in bars. A bear will surely be a disturbance in any bar, but the rule does not apply directly to bears. Rules are then based on an *entrenched generalization*: they are at once *under-inclusive* (the bears) and *over-inclusive* (the eye-seeing dogs). Standards are comparatively flexible: "Only good dogs are admitted". The goodness of a dog should be assessed according to the justification of the standard. Someone should consider what may disturb customers

and decide whether each particular dog may behave accordingly. This is the soft evaluative trigger. And only then, she may allow the dog in the bar.

There is a substantial difference between standards and rules in terms of costs. There are more *cognitive costs* to standards than to rules: these latter are simple and easy to apply; the former require interpretation and decision-makers should study all the relevant features of potential triggers, before authorizing the response. Both standards and rules have *patient costs* in case they lead to the wrong decision: in the case of pharmaceutical regulators, how often decisions based on either standards or rules grant market entrance to unsafe or ineffective drugs. Assuming a similar *error rate* for both, the more a directive is applied, the more rules are preferable to standards. There are less cognitive costs to rules: the comparative advantage of different triggers is studied and assessed only once and then implemented automatically.

Following the passage of the 1962 Food and Drug Administration Act, and up until 2016, drug regulation in the United States has been based on a combination of rules and standards. The 1962 Act established that drug manufacturers had to apply to the FDA for the approval of new treatments. Such application would only be considered if:

> *"[I]t 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" (Katz, 2004).*

The definition of a "well controlled investigation" was further clarified in 1969 (Section 314.126 of Title 21 of the Code of Federal Regulations), when it was formally quantified as two randomized controlled trials plus one previous or posterior confirmatory trial. Clinical trials are comparative experiments in which the effects of an experimental drug are compared to the standard treatment or a placebo, testing the hypothesis that there is no difference between them according to a frequentist statistical design. Once the manufacturer has collected all the necessary evidence, it is submitted to the FDA for review in the form of a New Drug Application (NDA).

FDA approval is then generally based on a rule: *if there are two positive RCTs, grant market authorization*. But there are also occasions to use standards. For instance, the reviewers might disagree on the quality of the evidence submitted. In such cases, the FDA may summon an advisory committee of experts to resolve the dispute (Pray and Robinson, 2007). Their conclusion is not mandatory for the FDA, although it is usually attended (Zuckerman 2006). The FDA directly appoints members of the advisory committee from experts in the relevant scientific fields. For example, the Oncologic Drug Committee consists of 13 voting members including one representative of patients, while the industry is represented with one non-voting member.

Usually the FDA and the company that submitted the NDA present all the relevant data to the members of the committee, who should then proceed to a public collective deliberation. Once this is closed, each member of the committee must vote on some specific yes-or-no questions formulated by the FDA, such as "Given the current knowledge, does this medicine have a demonstrated benefit?" (Urfalino 2012). The FDA itself stresses the fact that it pays attention not only to the votes, but also to the deliberation process (Urfalino 2012). The entire process is as transparent as possible: the meetings of the committees are open to the public, and the minutes are then published on the FDA website.

In general, the questions for new drugs concern "whether the safety and effectiveness information submitted for a new drug is adequate for marketing approval" (Center for Durg Evaluation and Research, n.d.). The members of these advisory committees should thus cast their votes according to a standard: *if the evidence presented is substantive enough, grant market approval*. This is clearly a soft evaluative trigger, as compared to the outcome of an RCT (e.g. either it reaches statistical significance or not) and each voter may interpret it differently. We should now discuss the costs involved in both regulatory approaches.

3.3 The cognitive costs of regulatory rules and standards

Historically there have been different ways to assess the safety and efficacy of new treatments. Between 1900 and 1950 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 considered 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 patients): 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. The following items counted as controls in these trials: the patients' eligibility criteria, the way treatments were allocated (alternation and randomization), uniformity in administration of treatments and patients' blinding. They were not used

necessarily all at once. Statistical reports from controlled trials conveyed their results with different degrees of sophistication. The standardization of what we now call an RCT began with the British Streptomycin trial in 1947 and concluded with the adoption of RCTs as a regulatory yardstick by the FDA in the 1960s. RCTs articulated the controlled experiment with the template statistical design first advocated by Ronald A. Fisher (see chapter 1).

The different methods for drug testing were all potential candidate triggers in a regulatory *rule*. As compared to these other methods, RCTs are considered superior at grasping causality. Regulatory agency should assess the effects of new treatments in terms of their safety and efficacy. Following Cartwright's analysis, RCTs can *clinch* these effects: "if the assumptions of the study design are met, a positive result deductively implies that the [intervention] under test causes the outcome under investigation in some study members" (Cartwright and Hardie 2015). In this respect, it is epistemically cheap to adopt them as triggers in our regulatory rules: we have an a priori argument to prefer RCTs to any of the above mentioned alternatives.

Nonetheless, we can interpret Cartwright's analysis as an equally a priori argument for the inevitability of standards in drug regulation. RCTs only prove, if positive, that the intervention had a given effect on the particular group of patients under study. Pharmaceutical regulators care instead for the safety and efficacy of a treatment in the whole population of potential patients they should protect. According to Cartwright, they should not assume that causes are simply necessary and sufficient conditions for producing their effect, as they currently do. For Cartwright, following Mackie (1974), causes are instead like cakes, with their many ingredients. A short-circuit is not a necessary and sufficient condition for a fire. There are many other contributing factors (the ingredients in the causal cake): e.g. the absence of sprinklers or the presence of oxygen. All these factors are called Insufficient but Necessary part of

an Unnecessary but Sufficient (INUS) conditions. A treatment effect is not just caused by the treatment alone, but in conjunction with a number of contributing factors. An RCT should ideally list (and control for) all these factors. When the treatment is applied outside the trial, it will only produce the same effects as long as the contributing factors are equally present. Physicians should assess "what the evidence has to say about the efficacy of an intervention for particular patients in a particular practice setting" (Fuller 2013).

Regulatory authorities should not blindly apply the rule *if there are two positive RCTs, grant market approval*. For Cartwright, they should find evidence about the causal factors that will contribute to the efficacy of the intervention on the general population of patients. There is no single best method for such a causal inference that we could use as a trigger in a regulatory decision rule. Cartwright suggests instead collecting all the facts that are relevant to the transition from efficacy in the trial to effectiveness on the general population. For this search there is no algorithmic method, but rather a number of heuristics that regulators should use wisely: e.g. they can conduct a pre-mortem analysis, a thought experiment of the form: "If the intervention goes wrong, how will it have gone wrong?". This will provide some clues as to the contributing factors that may spoil the effect of a treatment outside the trial. These heuristics cannot be synthesized in a simple rule with a hard empirical trigger. Drug regulation should inevitably use *standards* with soft evaluative triggers: *if the evidence presented is substantive enough, grant market approval*. Deciding about the substantiality of evidence requires the sort of deliberation carried out by the advisory committees at the FDA. Regulatory authorities cannot spare themselves the costs of deciding on a case-by-case basis what evidence is best for each given treatment. As the legendary FDA officer Robert Temple once put it (Berry et al. 2005), this case by case approach without clear decision rules was "his idea of a nightmare".

3.4 The costs of impartial deliberation

Is Temple's regulatory nightmare worth having? Ultimately, it depends on whether the increased cognitive costs of standard-based decisions are compensated with a lower error rate. If Cartwright is right, we may expect deliberative committees to perform better than rule-based review boards in detecting unsafe or inefficient drugs. *But we should be aware that deliberation per se does not seem to yield very different decisions*. A study conducted by the National Research Center for Women & Families provides a quantitative analysis of the decision of FDA advisory committee (Zuckerman 2006), following their voting patterns. One of the most interesting findings is that committee members voted unanimously for 66% of the drugs they recommended for the approval, which is particularly surprising given that advisory committees are convened for controversial cases in which the available RCT evidence does not seem particularly persuasive. Moreover, sometimes the FDA has approved drugs for which the advisory committees recommended the opposite. In Zuckerman's analysis, between 1998 and 2005 advisory committees voted against 11 drugs, but the FDA subsequently approved 4 (36%) of them. Whereas according to another report (Smith et al. 2012) the FDA decisions are very consistent with the advisory committee voting: in a set of 63 FDA advisory committee meetings that included votes for or against the approval of a new drug between 2001 and 2010 only 2 times the FDA approved a drug despite the committee voting.

According to Zuckerman "many of today's FDA drug and device advisory committees are rubber stamps for approval almost every time they meet" (Zuckerman 2006). We have no statistics about withdrawals and warnings for committee decisions, but there have been significant scandals on both accounts. For example, there were unanimous votes for the approval of Celebrex in 1998 and Vioxx in 1999, two drugs that subsequently were found to significantly increase the risk of heart attack and

stroke. Vioxx was subsequently withdrawn, giving rise to a public scandal, while Celebrex remained on the market, but with strong warnings.

In other words, despite its superior cognitive costs, standard-based deliberation on its own may lead to the same mistakes than rule-based approvals. There is a growing consensus about the problem not being in the decision method as such, be it standards or rules, but on the external pressure that the pharmaceutical industry imposes on regulatory bodies. For the industry, regulatory decisions are high-stakes bets: developing a drug to the point of submitting a NDA to the FDA is expensive, it does cost an average of $2.6 billion merely to get a drug through the FDA approval process (Mullard 2014). Thereby the temptation for the industry to "capture" the regulator and find ways to make the FDA decide according to the industry interest, even if it collides with the interests of the public. Critics of the pharmaceutical industry have extensively shown the different ways in which both rules and standards can be flouted, getting drugs approved without clear evidence of their safety and efficacy.

As to RCT based rules, there are indeed many contentious points in the design and conduct of trials: the list is too long to review it in full here. As of today, the sponsor of the trial decides about all these points and there is a large body of evidence showing that trials funded by industry are designed in a way that is likely to provide positive outcomes for the sponsor. Bero and Rennie (Bero and Rennie 1996) have discussed several ways in which industry-sponsored trials might favor the outcomes of a trials. For example, researchers could decide to use inappropriate controls. This means that the experimental drug might be tested against a placebo even when an efficacious therapy is already available, or use a dosing in the control group which favor the experimental drug. With regard to this, a study (Rochon et al. 1994) showed that almost half of industry sponsored of non-steroidal anti-inflammatory agents adopt arbitrary dosing which systematically favoring the experimental drugs over the

controls. Bias can also be introduced implementing inappropriate outcome measures. Surrogate endpoints are nowadays extensively used to accelerate drug development and save money, but their validity is questionable (Aronson 2005; Pereira, Horwitz, and Ioannidis 2012). Measuring an effect on a surrogate endpoint may suggest that the drug is very effective while it is not, or it can be found to have severe toxicity in the long-term. A well-known example is that of two antiarrhythmic drugs, encainide and flecainide which were shown "to decrease premature ventricular contractions after myocardial infarction" (Montaner, O'Shaughnessy, and Schechter 2001). But a later trial with "death" as primary endpoint found that those two drugs actually increased mortality (compared with placebo) (CAST Investigators 1989). So, in the end pharmaceutical companies sponsoring trials can make many methodological choices that deviate from the ideal study design for RCTs, and which *a priori* can favor the experimental drug.

But committees are equally sensitive to pharmaceutical pressure. According to many FDA critics, conflicts of interest are pervasive among committee members. Indeed many, if not all, of the committee members have some financial ties with one or more pharmaceutical companies. A recent study (Pham-Kanter 2014), analyzing the voting pattern from 379 CDER meetings during the 15-year period 1997-2011, showed a strong pro-sponsor bias among the committee members who have financial relationships with firms. Interestingly, this pro-sponsor bias appears to be larger when the scientific evidence is more ambiguous, such as in all those cases in which a deliberative judgment would be necessary to make a good decision.

Therefore, even if Cartwright is a priori right, and deliberative committees should allow us to make better regulatory decisions (with lower error rates than RCT-based rules), in order to see such decisions emerge we need to protect the experts from third-party influences. Cartwright remains silent as to this problem, but other

philosophers have actively tackled this problem. Let us just consider two different approaches: Kevin Elliott (Elliott 2016) more radically suggests eradicating the very source of those influences, while Justin Biddle (Biddle 2007, 2013) defends instead to make those influences explicit in the deliberative process in order to offset them. From our perspective, the key point is how both approaches increase the already high costs of standard-based deliberation.

According to Elliott, experts deliberating on evidence in order to make a standard-based decision may be most easily biased in the following circumstances (Elliott 2016):

(1) Scientific findings are ambiguous or require a good deal of interpretation or are difficult to establish in a straightforward manner.

(2) Individuals or institutions have strong incentives to influence those scientific findings in ways that damage the credibility of the research.

(3) Individuals or institutions that have incentives to influence those scientific findings also have adequate opportunities to influence them.

For Elliott, expert deliberation would be more impartial if these potential sources of bias were controlled for. Of course, controls of this sort have been already implemented with different degrees of success. Conflict of interest policies are a case in point: Intemann (Intemann and de Melo-Martín 2014) shows that disclosure policies do little to prevent or identify bias. When the bias is unconscious, disclosure policies will not reveal it. When the potential source of bias is acknowledged, and disclosed, reviewers and readers are not provided with any tools to identify the effects of those biases in the argument. Disclosure policies work at most as a red flag, with an unexpected side effect: if conflicts of interests are declared, readers tend to devaluate the research findings indiscriminately.

The cognitive costs of standards are already high (as compared to rules). Protecting committees from biases seems to raise these costs, as the disclosure policies example illustrates. And usually these cognitive costs have financial implications as well, although the resources available to regulatory agencies are often modest. For instance, funding for the FDA's Center for Drug Evaluation and Research for fiscal year 2007 was about $500 million (Institute of Medicine (US) Forum on Drug Discovery Development and Translation 2007) that is less than the average total cost of a large RCT. Moreover, about half of this budget comes from the user fees, i.e. fees payed by pharmaceutical companies for submitting to the FDA a New Drug Applications.

An alternative approach to warranting the impartiality of standard-based decision making is to make conflicts of interest explicit and try to strike some sort of balance, instead of just suppressing them. Drawing on Kantrowitz (Kantrowitz 1967, 1976) and Merton (Merton 1973), Biddle has argued for the introduction of an adversarial system in regulatory committees, in which the interests in conflict are explicitly argued for: “Two groups of advocates would present arguments for a specific position, and a panel of judges would adjudicate between these two groups” (Biddle 2013). The industry chooses its own advocates, whereas the other set of advocates can be chosen among industry competitors (insurance companies, public health care agencies, patients and citizen groups). Judges of course must be scientific experts independent from both parties, having no direct interest or connections to the issue under consideration. The advocates and the judges should be, of course, kept separate. To control for independency, both groups of advocates should have the possibility to exclude a given number of scientists from the panel of judges. In this way, the decision should be as much impartial and objective as possible.

Rather than investing a massive amount of resources in creating fully impartial committees, Biddle’s suggestion is to invest a limited amount in creating a body of

expert judges without conflicts of interest and let the advocates absorb the costs of finding the right evidence for their interpretation of the standard. We should be nonetheless aware of the inequality in resources between the contending parties. Most patients' organizations are poor and pharmaceutical companies are their primary source of funding. To illustrate this, consider that in the notorious Vioxx scandal, Merck set up a $4.85 billion settlement fund just to resolve consumer claims, while in 2014 the total revenues of one of the largest patient advocacy organization (the American Heart Association) were just 774 million (Forbes)[1].

In other words, assuming Cartwright's arguments about the superiority of deliberative committees deciding upon standards, in the case of regulatory decisions, we should add to the cognitive costs of finding the right evidence the additional costs of protecting committees from third party influences. And these cognitive costs usually translate into more resources. When the third party is as financially powerful as the pharmaceutical industry, we cannot expect to obtain cheap and good standard-based decisions.

3.5 An experimental approach to regulatory decision-making

Philosophers, like Cartwright, Elliott and Biddle, advocate for reforms of our regulatory procedures on a priori grounds. For Cartwright, the cognitive costs of deliberating on standards are worth for their inferior error rate (as compared to rule-based decisions). Assuming the epistemic superiority of deliberative committees, Elliott and Biddle justify the further costs of warranting their impartiality. A priori, it is indeed plausible that such committees may protect patients better. But in assessing the quality of

[1] "American Heart Association on the Forbes. The 100 Largest U.S. Charities List". Forbes. Accessed October 19, 2017. https://www.forbes.com/companies/american-heart-association/

regulatory bodies, we should take into account the cognitive and economic resources a society is willing to invest in them.

Even if we had impartial experts with all the available evidence at hand, making the right decision may not be within their reach. According to Peter Gøtzsche (Gøtzsche 2013), the clinical documentation for just three NDA can take up 70 meters of binders. Finding safety and efficacy issues in such a data deluge and making the right decision may be simply too costly, independently of whether rules or standards are used. There is indeed the possibility that, for the amount of resources the FDA has, patients are reasonably well protected with RCT-based rules. Or, at least, it is not self-evident than with a similar amount of resources standard-based decisions will protect patients better. We need to find an empirical index that allows us to judge which decision procedure is more suitable for protecting patients.

Let us assume that market withdrawals are a rough indicator of the error rate of RCT-based rules. When drugs have serious and frequent adverse effects, the FDA will study whether and how the drug's benefit and risk balance compares with treatment alternatives. If the risks outweigh the benefits, the drug is withdrawn from the market. Given that the FDA mandate is to test drugs for safety and efficacy, drug withdrawals are explicit regulatory mistakes. Following (Carpenter, Zucker, and Avorn 2008), there were only 4 safety-based withdrawals for the 216 drugs approved following the customary FDA procedure in between 1993 and 2004. This is a 1.9% error rate[2]. Critics of the FDA object that these figures are artificially low. Many dangerous compounds remain in the market with a *black box warning*: a note in the drug's label calls attention to serious or life-threatening risks. Following still Carpenter, there were 4 such warnings (1.9%) for drugs approved through the customary procedure.

[2] Interestingly, calculating the type II error rate (safe and efficacious drugs that were incorrectly rejected) is out of our reach, since information about not-yet approved NDAs is considered proprietary under FDA regulations (Carpenter, 2014)

Before discussing whether these error rates are reasonable, we should wonder whether they are reliable at all. Had the FDA been captured by industry interests, as many critics contend, there would be no reason to trust these figures: the real number of dangerous compounds in the market would be much higher. Again, we rely on Carpenter for the best discussion available of this problem. Regulatory capture occurs when industry biases regulatory decisions according to its particular interests, even if these latter conflict with the public interest that the regulator should protect. *Corrosive capture* occurs when an industry pushes the regulatory process "with the aim of reducing costly rules and enforcement actions that reduce firm profits", for instance, withdrawing drugs from pharmaceutical markets (Carpenter and Moss 2014). Successful corrosive capture implies that regulators do not comply with their statutory obligations in protecting consumers from unsafe drugs. Critics of the FDA contend that this is precisely what happened:

Scientists at drug agencies are not only up against a powerful industry, they are also often up against their own superiors and their advisory committees who may have less than ideal motives for their decisions. The bosses often look the other way because they depend on licensing fees and political goodwill, and because questions about harms lead to trouble. A culture develops where many decisions are made that ordinary citizens would not have agreed with if they had been represented in the drug advisory committees

A case in point of corrosive capture would be the 1992 Prescription Drug User Fee Act (PDUFA): companies were taxed for their NDA submissions in return for a within-deadline approval process. This clearly is in the financial interest of the industry: the sooner the drug is on the market, the more patent years can be commercially exploited. Drugs approved under the PDUFA between 1993 and 2004 track have a higher percentage of withdrawals (7%) and black box warnings (9%), as

compared to standard approvals. The difference is statistically significant as compared to the figures above (Carpenter, Zucker, and Avorn 2008). Assuming that the public interest is best served by a lower error rate, there would be here a deviation. However, Carpenter warns about the complexity of the case. On the one hand, the public interest may be broader in scope: behind the PDUFA we find not just the industry, but patient advocacy groups requesting quicker review processes in order to have earlier access to drugs; and the US Senate passed it unanimously. It remains an open question what exact trade-off between speed and safety is acceptable for pharmaceutical consumers. On the other hand, capture implies that this is not an unintended shift: there should be a clear connection between the passage of the PDUFA and this outcome. According to Carpenter, advocates of the capture interpretation have failed to establish this connection, beyond testimonies and anecdotal evidence. It would be necessary to show, for instance, that officers behind the PDUFA approvals have a statistically significant higher rate of conflicts of interest than those behind the standard approvals. Otherwise, it would be difficult to rule out the alternative explanation that is the deadline itself, rather than the speed of approval: with well-staffed teams, there is evidence that approvals may be equally fast, on average, without increasing safety risks (Carpenter, Zucker, and Avorn 2008). Hence, the increased error rate may be an unintended consequence of a misguided legislation (approval deadlines) for which there was consensual bipartisan support (and not just industry interests).

Summing up, following Carpenter's analyses, our claim is that market withdrawals (and black box warnings) provide an empirical benchmark of how well the FDA can meet its regulatory mandate on a rule-based system, with its current institutional arrangements. Critics of the RCT-rule should discuss under which circumstances a standard-based system can overperform such error-rate. If the FDA can be persuaded to run a pilot program, perhaps we will be able to find out empirically

which system is better. However, as a note of caution, in drafting this pilot program, it will be necessary to discuss the cognitive and financial costs of the improvement: how much more evidence they would need and how much protection from conflicts of interests. The open normative question is whether a lower error-rate is going to be worth those resources.

3.6 Conclusion

The distinction between rules versus standards allows us to capture the costs at stake in the philosophical controversies on regulatory evidence. There are, indeed, good a priori arguments about the superiority of standards of evidence in the detection of safety and efficacy. And there are equally good a priori proposals on how to organize impartial committees to carry out such deliberation. Framing the decision in terms of the costs of implementing these ideals allows us to grasp the actual trade-off that regulatory authorities should make: there is only a limited amount of resources for them to use. It is possible that RCT-based rules have offered so far a cost-effective way of protecting patients, if their safety is defined in terms of market withdrawals. In any case, our regulatory experience provides an empirical benchmark for other alternatives. We suggest that, rather than further a priori discussions, we need instead to test these alternatives in a pilot committee. As it has often happened in the past, philosophy has been good at setting up ideals, but philosophers of science may find that reality does not live up to them – as it has happened to moral and political philosophers in the past.

Our analysis is grounded in a simple and obvious fact: generally speaking, drugs always come with some sort of side effects, which eventually can harm the consumers. That is why the assessment of drug safety is a matter of a risk-benefit analysis. However, some critics of pharmaceutical system (e.g. Vandenbroucke 2004, Osimani

2014; Stegenga 2016) hold that since current standard methodologies – namely, RCTs – systematically underestimate harms, then any benefit-analysis would be biased. For example, according to Stegenga (2016) clinical research does not reliably measure the harms of drugs because every phase of trials suffers from different bias: publication bias, sensitivity issues, evidence shrouded in secrecy, etc. Also, clinical research would be biased upstream by the way harms are defined and measured. Vandenbroucke (2004, 2008) for its part, endorses the adoption of different methodological standards to better capture harms on the basis of a priori epistemological considerations. The "hierarchy of evidence" should be reversed when we are interested in finding unintended events. From a methodological point of view, observational/epidemiological studies would not suffer from selection bias: "Even if the doctor knows whom s/he is prescribing the treatment to, treatment allocation is masked with respect to unintended effects, given that s/he does not know them" (Osimani 2014). From an epistemological point of view, with respect to adverse effects, we aim to *discover* harms rather than *test* a hypothesis of efficacy (or safety). And in the context of discovery, other methodologies, such as case reports, can provide more convincing evidence than RCTs. These criticisms have two main implications for our discussion.

First, drugs withdrawals for safety reasons would be a good empirical benchmark only if they were based on evidence coming from sources different than RCTs. This is exactly the case, at least according to a recent systematic review on worldwide withdrawal of medical products because of adverse drug reactions (Onakpoya, Heneghan, and Aronson 2016). As a matter of fact, case reports are cited as evidence in 30 out of 40 withdrawals. Moreover, the entire surveillance system looks quite efficient, considering that the median interval between the first report of an adverse event and the withdrawal of the correspondent drug is 1 year, and the time it

takes for drugs to be withdrawn is shortening over years. Also, the risk-benefit is assessed on an orthodox interpretation of the precautionary principle, when the evidence is in doubt withdrawal is more likely, especially when dramatic effects (e.g. deaths) are reported.

Secondly, since standard definitions of harms are too under-inclusive we cannot have a reliable estimate of the safety of any drug. It is true that RCTs are not designed to capture all the potential harms of a drug. According to Gøtzsche (2013), for instance, drugs are usually tested in a different population than the one who will actually take them in the real world. For example, the largest consumers are elderly, who are systematically excluded from clinical trials. Furthermore, elderly people are usually in treatment with several drugs, and we know very little about polypharmacy. But an ideal system collecting all the potential evidence about the safety of a drug is likely outside the resources of regulators. Anyway, the additional cognitive and financial costs should be discussed in connection with the real benefits for the patients. As we stated above, the open question is whether such an ideal system is really worth it.

With regard to this, the last point we want to make is that benefit-risk analysis should not be assessed *prima facie* from a public health perspective and it is not only a matter of quantifying the adverse effects of a drug. The idea is that knowing in advance all the harms (assuming that this is feasible) would not impact much at the level of individual patients. Ultimately, protecting patients depends more on the prescribing physicians and the risk-benefit balance for the individuals rather than on the detection of harms at the population level. But that analysis also depends on the amount of risk that one person is willing to accept, and the variability of risk-aversion is very high among individuals. According to Edwards (Edwards 2005) there are very few drugs in which the overall risk really is greater than the effectiveness. Thus, when a drug is taken off the market, all the patients who were taking the drugs without any adverse

effects are going to suffer. Therefore, Edwards suggests focusing more on how patients are actually treated rather than whether a drug is on the market or not. Patients should be asked what risks they are prepared to take and for what benefits. From the individual perspective Vioxx scandal is not such a big failure of drug regulation itself. In the debate, patients who needed Vioxx (or any COX 2 inhibitors) have been left without any alternative: some of them have had only a disadvantage from Vioxx withdrawal. Therefore, a more reliable benefit-risk analysis would require much more useful information coming from different studies addressing different safety and efficacy questions, such as for instance a comparison between Vioxx and older NSAIDs. And yet the usefulness of information is hardly identifiable *a priori*.

References

Aronson, J. K. 2005. "Biomarkers and Surrogate Endpoints." *British Journal of Clinical Pharmacology* 59 (5):491–94. https://doi.org/10.1111/j.1365-2125.2005.02435.x.

Bero, Lisa A., and Drummond Rennie. 1996. "Influences on the Quality of Published Drug Studies." *International Journal of Technology Assessment in Health Care* 12 (2):209–37. https://doi.org/10.1017/S0266462300009582.

Berry, Donald A, Steven N Goodman, and Thomas A Louis. 2005. "Floor Discussion." *Clinical Trials* 2 (4):301–4. https://doi.org/10.1191/1740774505cn101oa.

Biddle, Justin. 2007. "Lessons from the Vioxx Debacle: What the Privatization of Science Can Teach Us About Social Epistemology." *Social Epistemology* 21 (1):21–39. https://doi.org/10.1080/02691720601125472.

———. 2013. "Institutionalizing Dissent: A Proposal for an Adversarial System of Pharmaceutical Research." *Kennedy Institute of Ethics Journal* 23 (4):325–53. https://doi.org/10.1353/ken.2013.0013.

Carpenter, Daniel, and David A. Moss. 2013. *Preventing Regulatory Capture: Special Interest Influence and How to Limit It*. Cambridge University Press.

Carpenter, Daniel, Evan James Zucker, and Jerry Avorn. 2008. "Drug-Review Deadlines and Safety Problems." *New England Journal of Medicine* 358 (13):1354–61. https://doi.org/10.1056/NEJMsa0706341.

Cartwright, Nancy, and Jeremy Hardy. 2012. *Evidence-Based Policy: A Practical Guide to Doing It Better*. New York, USA: Oxford University Press USA. http://global.oup.com/?cc=gb.

"Collective Wisdom: Principles and Mechanisms." 2012. Cambridge Core. July 2012. https://doi.org/10.1017/CBO9780511846427.

Edwards, I Ralph. 2005. "What Are the Real Lessons from Vioxx?" *Drug Safety* 28 (8):651–58. https://doi.org/10.2165/00002018-200528080-00001.

Elliott, Kevin C. 2016. "Standardized Study Designs, Value Judgments, and Financial Conflicts of Interest in Research." *Perspectives on Science* 24 (5):529–51. https://doi.org/10.1162/POSC_a_00222.

Gøtzsche, Peter C. 2013. *Deadly Medicines and Organised Crime: How Big Pharma Has Corrupted Healthcare*. Radcliffe Publishing.

Intemann, Kristen, and Inmaculada de Melo-Martín. 2014. "Addressing Problems in Profit-Driven Research: How Can Feminist Conceptions of Objectivity Help?" *European Journal for Philosophy of Science* 4 (2):135–51. https://doi.org/10.1007/s13194-013-0079-9.

Kantrowitz, Arthur. 1967. "Proposal for an Institution for Scientific Judgment." *Science* 156 (3776):763–64. https://doi.org/10.1126/science.156.3776.763.

Katz, Russell. 2004. "FDA: Evidentiary Standards for Drug Development and Approval." *NeuroRX* 1 (3):307–16. https://doi.org/10.1602/neurorx.1.3.307.

Mackie, John L. 1974. "The Cement Ofthe Universe." *London: Oxford Uni*.

Merton, Robert K. 1973. *The Sociology of Science: Theoretical and Empirical Investigations*. University of Chicago press.

Montaner, Julio SG, Michael V O'Shaughnessy, and Martin T Schechter. 2001. "Industry-Sponsored Clinical Research: A Double-Edged Sword." *The Lancet* 358 (9296):1893–95. https://doi.org/10.1016/S0140-6736(01)06891-X.

Mullard, Asher. 2014. "New Drugs Cost US$2.6 Billion to Develop." *Nature Reviews Drug Discovery* 13 (12):nrd4507. https://doi.org/10.1038/nrd4507.

Onakpoya, Igho J., Carl J. Heneghan, and Jeffrey K. Aronson. 2016. "Worldwide Withdrawal of Medicinal Products Because of Adverse Drug Reactions: A Systematic Review and Analysis." *Critical Reviews in Toxicology* 46 (6):477–89. https://doi.org/10.3109/10408444.2016.1149452.

Osimani, Barbara. 2014. "Hunting Side Effects and Explaining Them: Should We Reverse Evidence Hierarchies Upside Down?" *Topoi* 33 (2):295–312. https://doi.org/10.1007/s11245-013-9194-7.

Pereira, Tiago V., Ralph I. Horwitz, and John P. A. Ioannidis. 2012. "Empirical Evaluation of Very Large Treatment Effects of Medical Interventions." *JAMA* 308 (16):1676–84. https://doi.org/10.1001/jama.2012.13444.

Pham-Kanter, Genevieve. 2014. "Revisiting Financial Conflicts of Interest in FDA Advisory Committees." *Milbank Quarterly* 92 (3):446–70. https://doi.org/10.1111/1468-0009.12073.

Pray, Leslie, and Sally Robinson. 2007. "Challenges for the FDA: The Future of Drug Safety." In *Workshop Summary. National Academies*.

Research, Center for Drug Evaluation and. n.d. "Investigational New Drug (IND) Application - Drug Development and Review Definitions." WebContent. Accessed October 24, 2017. https://www.fda.gov/drugs/developmentapprovalprocess/howdrugsaredevelopedandapproved/approvalapplications/investigationalnewdrugindapplication/ucm176522.htm.

Rochon, Paula A. 1994. "A Study of Manufacturer-Supported Trials of Nonsteroidal Anti-Inflammatory Drugs in the Treatment of Arthritis." *Archives of Internal Medicine* 154 (2):157. https://doi.org/10.1001/archinte.1994.00420020059007.

Schauer, Frederick. 1991. *Playing by the Rules: A Philosophical Examination of Rule-Based Decision-Making in Law and in Life*. Clarendon Press.

Schlag, Pierre. 1985. “Rules and Standards.” *UCLA Law Review* 33:379.

Science, American Association for the Advancement of. 1976. “The Science Court Experiment: An Interim Report.” *Science* 193 (4254):653–56. https://doi.org/10.1126/science.193.4254.653.

Smith, Jeffrey F., Seth A. Townsend, Navjot Singh, and Philip Ma. 2012. “FDA Advisory Committee Meeting Outcomes.” *Nature Reviews Drug Discovery* 11 (7):nrd3747. https://doi.org/10.1038/nrd3747.

Stegenga, Jacob. 2016. *Measuring Harms*. Routledge Handbooks Online. https://doi.org/10.4324/9781315720739.ch31.

The Cardiac Arrhythmia Suppression Trial (CAST) Investigators. 1989. “Preliminary Report: Effect of Encainide and Flecainide on Mortality in a Randomized Trial of Arrhythmia Suppression after Myocardial Infarction.” *New England Journal of Medicine* 321 (6):406–12. https://doi.org/10.1056/NEJM198908103210629.

U. S. Government Accountability Office. 1994. “FDA User Fees: Current Measures Not Sufficient for Evaluating Effect on Public Health,” no. PEMD-94-26 (August). http://www.gao.gov/products/PEMD-94-26.

Vandenbroucke, Jan P. 2004. “When Are Observational Studies as Credible as Randomised Trials?” *The Lancet* 363 (9422):1728–31. https://doi.org/10.1016/S0140-6736(04)16261-2.

Vandenbroucke, Jan P. 2008. “Observational Research, Randomised Trials, and Two Views of Medical Science.” *PLOS Medicine* 5 (3):e67. https://doi.org/10.1371/journal.pmed.0050067.

Diana M. Zuckerman 2006 FDA Advisory Committees: Does Approval Mean Safety? Washington, DC: National Research Center for Women & Families.

4. DRUG REGULATION AND EVIDENTIARY PLURALISM

"If you tried to doubt everything you would not get as far as doubting anything. The game of doubting itself presupposes certainty"

(Ludwig Wittgenstein, On Certainty)

4.1 Introduction

In early December 2016, the Congress of the United States passed the 21st Century Cures Act (21CCA), the biggest health reform legislation since Obamacare. The bill is vast, ranging from provisions related to the funding and administration of the National Institute of Health, to new rules for informed consent, clinical data sharing, and patients' privacy. The sections aimed at getting new drugs and medical devices approved more easily and quickly has been the most controversial so far. The 21CCA (Public Law 114-255) introduces a more flexible drug approval process, with less demanding requirements for certain drugs and new evidentiary standards as an alternative to the expensive and time-consuming RCTs. For instance, in order to support the approval of certain drugs for new indications, the 21CCA allows pharmaceutical companies to provide "data summaries" and "real world evidence" (e.g. observational studies), instead of RCT data (Sec. 3022). Moreover, the law gives more flexibility to the FDA to grant accelerated approval to new drugs for life-threating conditions or unmet medical needs, on the basis of early-phase clinical trials, trials measuring efficacy and safety on surrogate outcomes alone, or trials adopting heterodox experimental designs (Sec. 3021). Critics of the 21CCA argue that the new

legislation is a "gift" to pharmaceutical companies, its most active promoters, with a large amount of expenses in lobbying activities[1]. In particular, critics feel that shifting away from RCTs will open the door to harmful or ineffective treatments. In the end, patients will suffer the consequences. Nonetheless, patients' advocacy groups actively supported the 21CCA, mainly because it promises faster market access to new treatments (Avorn and Kesselheim 2015; Zuckerman, Jury, and Silcox 2015).

The case of solanezumab illustrates the dilemma. While the 21CCA was under discussion at the US Senate, the Food and Drug Administration (FDA) rejected solanezumab, a drug for Alzheimer keenly awaited by many patients, since it showed promising results in early phase trials. Yet, it turned out to be no better than a placebo in standard RCTs. Kesselheim and Avorn speculate on what might have happened if a more flexible evidentiary standard had been already in place. Since there are no working treatments for Alzheimer, early phase trials suggesting benefit for patients with Alzheimer's disease would have "been enough to support FDA approval — particularly since no major safety concerns had been raised in those trials" (Kesselheim and Avorn 2017). As Kesselheim puts it, the 21CCA changes in the approval process "are based on the foundational misconceptions that the FDA standards for approval are too demanding and thus keep valuable new treatments from the US public and needlessly increase the cost and duration of drug and device development. In this respect, the law is a solution to a problem that mostly does not exist" (*ibi*). And yet, according to a letter from a US senator, the FDA is staring at about 4,000 new drug applications that have yet to be approved.

[1] The 21st Century Cures Act is one of the most lobbied bill of legislation in recent history. According to the Center for Responsive Politics, more than 1,455 lobbyists representing 400 companies have made their case for, or against, the law. Data on lobbying activities are free available at https://www.opensecrets.org/lobby/billsum.php?id=hr6-114. Accessed 19, October 2017.

According to its critics, the problem with the 21CCA is that it introduces not just a double, but a multiplicity of evidentiary standards for new regulatory approvals: whereas treatments undergoing the traditional testing process (conventional RCTs) stick to the old pre-21CCA standards of safety and efficacy, those other treatments (successfully) tested with new methods will bring into the market potentially different thresholds of safety and efficacy. Patients may end up suffering the consequences if these new methods fail. In this chapter, we want to argue that the multiplicity of testing standards is more defensible than critics think. As a matter of fact, since 1962 we already have a variety of standards for testing the safety and the efficacy of medical treatments, and the system has worked reasonably well so far (see chapter 3). As we will argue in section two, medical treatments (not just drugs) have different testing standards, according to the potential public health risks they pose. Drug regulation has developed, throughout the 20th century in reaction to public health catastrophes caused by all sorts of medical interventions.

Although we lack a principled agreement on what constitutes one such public catastrophe (e.g. how many victims? What age? Etc), the pre-21CCA regulatory schemes are grounded on an implicit consensus on the risks involved by different medical interventions: the bigger the threat, the stricter the testing standards. We shall illustrate it in sections 3 and 4 with the cases of surgery and medical devices. In section 5, we will present a concept of risk that, in our view, captures our current regulatory consensus. Risks depend on two factors: the hazards involved in a treatment and the number of people potentially exposed to it. Again, we lack a principled definition of both factors, but our current regulation has consensual enough approximations to both of them. From a political standpoint, this concept of risk is all we need to justify the existence of multiple testing standards and if it has worked well so far, there is no reason to expect its collapse with the 21CCA. Consequently, different testing standards

are defensible for certain class of drugs (e.g. targeted therapies). To conclude, in section 6, we will address a potential crucial objection to our argument: that any a priori assessment of risks always comes with an unacceptable degree of uncertainty, therefore we should always demand for stricter testing standards.

4.2 What the FDA does: the pervasiveness of double standards

Since the 1962 FDA Act, pharmaceutical regulatory agencies all over the world have required evidence of safety and efficacy before granting market approval. The United States FDA is the largest and the most influential of all of these agencies: it monitors a market of products worth over 1 trillion dollars, which represent nearly a fourth of consumers' spending[2]. With respect to medical products regulation, the FDA is generally considered a successful agency as far as we consider drug withdrawals as a reliable empirical benchmark. For instance, between 1993 and 2006 only a mean of 1.5 drugs per year have been withdrawn, and the number of withdrawals has not increased over time (Issa et al. 2007), moreover the agency performed well in protecting patients from some big failure (see Sacks, Avorn, and Kesselheim 2017).

The first point in our argument is that, well before the 21CCA, the FDA already operated under a multiple testing standard regimen[3]. The FDA classifies the medical products it regulates under three main categories according to their "nature", each of them provided with its own testing standard. If the product is a technological manufacture, it is considered a *medical device*; if the product is composed of biological compounds (sugar, proteins, nucleic acids, etc.), it is a *biologic*; finally, if the product is chemically synthesized, it falls on the category of *drugs*. Regulatory guidelines vary between the three categories. Drugs are tested through rigid controlled trials, while

[2] See https://www.fda.gov/AboutFDA/Transparency/Basics/ucm553038.htm. Accessed 19, October 2017.
[3] See Title 21 Code of Federal Regulations (21CFR) Parts 800 – 1299.

medical devices and biologics testing standards differ according to whether they are intended as treatments or not, ranging from basic manufacturing control to RCTs.

Why this classification and the different testing regimes? Intuitively it should be based on the different "nature" of the tested treatments, but this is less obvious than it may seem. Some products now defined as medical devices were previously classified as drugs. Biologics can be defined either as drugs or devices and therefore they are also subject to different regulations. Stem cells therapy is a non-trivial example, in 2006 the FDA extended its regulatory power over cell therapies, which before were considered as part of medical practice (as a common procedure of transplantation) and therefore not subject to any regulatory burden.

If the different testing regimes are not clearly grounded on the nature of the treatment examined, how can it be justified? The goal of the FDA is to protect consumers from harms (dangerous treatments) and scams (fake treatments) and yet, depending on the treatment, the FDA may apply different methods with potentially different consequences (just as critics of the 21CCA fear today). Of course, there is a clear difference in scientific rigor between RCTs and basic control of manufacturing: clinical trials are more severe than any other testing standard. Although, intuitively for some medical devices, biologics and drugs can have the same potential bad consequences for patients. That is precisely the reason why the FDA extended its control over stem-cells therapies (Freeman and Fuerst 2012), considering them as drugs and therefore assigning them to most rigid standards of testing: without any control, they can have terrible consequences, also considering that in theory they might be employed for a wide range of conditions. At the same time, other medical products (devices or biologics) seem to be less harmful for patients, or harboring different risks: rubber gloves if good-manufactured will hardly harm patients. Indeed, historically there have been medical catastrophes in pharmacology, but not so much elsewhere.

The FDA can wield its authority according to the potential risks involved in medical products. As well, the choice of testing standards (and of consequently regulatory burden) varies according to the risks. The two following sections will better illustrate this point.

4.3 What the FDA does not: the case of surgery

The existence of a double standard in testing medical treatments has been detected, debated and criticized (Deyo 2004), as we will now illustrate with the case of surgery. As we will show, there are epistemic and political reasons against the conduct of RCTs in surgery, but in our view only the latter explain the existence of a double standard for surgical treatments, and a very poor one at that. But this is a test case for our claim as well: it is not just that there have been political arguments for not regulating surgical procedures, there has been no positive reason to do it either. Our societies are apparently willing to live with the level of risks involved by our testing standards in surgery.

Let us review the arguments against conducting RCTs in surgery. First of all, on the epistemic side, some key features of RCTs cannot be easily implemented in surgical trials. Blinding and placebos are particularly difficult: surgeons cannot be blinded in any way, whereas the adoption of sham procedures as controls raise difficult ethical issues. Without proper blinding, biases can contaminate the trial outcome making the experiment scientifically useless. Patients are also reluctant to undergo treatment randomization in surgery. The procedures under test are often perceived as highly unequal, regarding the invasiveness of the intervention, side effects, or quality of life. Hence patient enrollment becomes difficult, making trials difficult to complete for lack of a proper sample size.

Surgical treatments are often difficult to standardize as RCTs require. Surgical interventions are complex and heavily skill-dependent procedures, subject to improvements in technical performance. Quality in performance requires extensive training over time – a "learning curve", which is difficult to control in a clinical trial. During the learning curve phase, errors may likely occur, and this could greatly distort the outcome results.

Difficult as these problems may be, there are alternative approaches that would make surgical trials feasible (Lilford et al. 2004; McCulloch et al. 2002). For instance, control techniques such as audit data collection or continuous third-party monitoring, may be included into the trial design to mitigate the lack of blinding and randomization. Learning curves and variation in technique can be measured and controlled as well (Farrokhyar et al. 2010). Also, according to the least optimistic review (Solomon and McLeod 1995) in the ideal situation RCTs can be performed to evaluate 40% of treatment questions involving surgical procedures – and the authors consider it a conservative estimate.

On the political side, there are two major obstacles for the conduct of surgical trials. On the one hand, there is the funding issue: RCTs are very expensive experiments and, unlike with pharmacological treatments, there is no clear sponsor for surgical tests. Publicly funded surgical trials are not a priority for any political party, as of today. On the other hand, surgical interventions are considered, at least in the US, a medical practice, outside the scope of any dedicated regulatory body: the FDA can only examine products, not services. In the 1960s and 1970s, the US medical profession resisted State intervention on medical practice (e.g. prescription) as a form of *communism* (see Tobbell 2011). For the same reasons, the FDA do not regulate behavioral and psychological interventions, nor off-label use of drugs.

In our view, the double standard in surgery owes more to politics than to method. The current testing procedure, on its own, seems difficult to defend: surgical treatments are regularly applied without any experimental testing, relying mostly on experts' judgment, anecdotal evidence, or case series. And these assessments often fail and surgical interventions are rarely riskless: as an example, extracranial-intracranial vascular bypass to reduce the risk of ischemic stroke has been widely performed based on a single case report. In 1985, however, a randomized trial demonstrated that extracranial-intracranial bypass actually increased the risk of fatal and nonfatal stroke compared to medical therapy alone, and the procedure has been abandoned ever since (The EC/IC Bypass Study 1985). Even if a double standard is inevitable, there is room for improvement in surgery (e.g. Tonelli, Benditt, and Albert 1996). Indeed, we may wonder why it is that so systematic attempt at reforming surgical testing has taken place.

Despite the prestige of the surgical profession, it is hard to believe that the community of surgeons is powerful enough to resist the external imposition of testing standards, when the omnipotent pharmaceutical companies are so heavily regulated. Here is our claim: if pharmaceutical regulation has been driven by a series of pharmacological catastrophes, there has been no comparable crisis in surgery to prompt regulators to act. The adoption of some inefficient and unsafe surgical procedures did not harm enough people neither to make regulators to demand for a special control over surgical treatments, nor for the politics to feel the necessity to invest resources to guard public health.

Medical catastrophes are unlikely to happen in surgery, that is mostly because of contingent reasons: new surgical techniques are usually successive adaptations of existing techniques lead to the emergence of new procedures that are not radical innovations produced by a specific research program, but part of a continuum formed

by the evolution of day-today practices (Frader and Caniano 1998). Occasionally new procedures arise in dramatic circumstances when surgeons, often in an emergency, decide to try a new approach even though there is no adequate statistical support for its efficacy. If they are successful, their techniques may subsequently form the basis of new protocols and be routinely applied (Petrini 2013). As soon as something goes wrong surgeons can easily dismiss a procedure, this means that in case of failure very few patients will suffer.

Societies seem to tolerate the risks associated with surgical procedures, leaving them unregulated. They do not seem a threat for public health, hence it is not worth to invest financial and cognitive resources to regulate them. As we are going to see in the next paragraph, the tolerance of certain risks associated with medical treatments is even clearer in medical device regulation.

4.4 The lessons from medical devices regulation

The regulation of medical devices provides additional evidence for our claim: regulators react to catastrophes alerting of the risks involved in certain treatments. And, as we are going to see, an a priori assessment of the potential risks of a treatment dictates the necessary level of testing.

The Thalidomide of medical devices' arena was the Dalkon Shield, an intrauterine contraceptive device marketed in 1971 by the A.H. Robins pharmaceutical company. With a massive advertising campaign, promising perfect birth control protection with virtually no adverse effects, it was a resounding success –the feared long terms side effects of contraceptive pills played a role here. Dalkon Shield was prescribed to more than 2 million women in the U.S, 10% of the market for contraceptives. In its three years on the market it caused more than 200.000 serious pelvic infections, leading to further adverse events ranging from infertility to death. In

the aftermath of the scandal, the Congress passed the landmark Medical Device Amendments (MDA) to the federal Food, Drug, and Cosmetic Act (FDCA), requiring extensive testing and formal regulatory approval before medical devices can reach the market.

The way the FDA handled this new mandate is equally illustrating for our claim. In 1976 the FDA suddenly had to sift through a huge volume of "medical products" falling under its jurisdiction. With scarce resources, the FDA had to prioritize its efforts. The FDA decided then to assign categories of risk associated with the devices, grounding on background knowledge and common sense. Intuitively, some medical products seem less risky than others: for instance, latex gloves naturally carry less risk than cardiac pacemakers, even if both count as devices for regulatory purposes. Hence, the agency created three different device classes, according to the hazards they pose to the patients, each class with its own testing standards. According to section 513(a)(1) of the FD&C Act (21 U.S.C. 360c(a)(1)), the three device classes are defined as follows:

- Class I: Devices are subject to a comprehensive set of regulatory authorities called general controls that are applicable to all classes of devices.
- Class II: Devices for which general controls, by themselves, are insufficient to provide reasonable assurance of the safety and effectiveness of the device, and for which there is sufficient information to establish special controls to provide such assurance.
- Class III: Devices for which general controls, by themselves, are insufficient and for which there is insufficient information to establish special controls to provide reasonable assurance of the safety and effectiveness of the device. Class III devices typically require premarket approval.

Class I devices are considered low-risk and do not undergo any regulatory review, they are subject only to control for good manufacturing practices. Class II

devices instead are believed to be moderate-risk, so they are subjected to a review procedure known as "510k" (from the section Section 510(k) of the Food, Drug and Cosmetic Act). In a 510k submission, the FDA establishes that a device is safe and effective if the manufacturer demonstrates that the new device is "substantially equivalent" either to an existing device already approved through the same process or to a device which has been on the market before the 1976 federal law was passed. These reference devices are named "predicate devices". Class III devices instead are high-risk, and/or have novel intended uses, and hence require direct demonstration of safety and efficacy through a process very similar to a New Drug Application (NDA): sponsor must submit valid scientific evidence of safety and efficacy. However, the standard of evidence is substantially lower for medical devices than for drugs.

While for drugs the FDA required "substantial evidence of safety and efficacy", specified in two positive RCTs, for medical devices it requires only "reasonable assurance of safety and efficacy" (Ciani, Tarricone, and Taylor 2016). This means, in practice, one single controlled study, but the regulation permits also "reliance upon other valid evidence […] even in the absence of well-controlled investigations" (*ibi*.). Thus, the evidence that would be never sufficient to grant market access to a drug can actually result in the approval of a medical device, maybe even to treat the same condition.

The regulation of medical devices makes the multiplicity of evidentiary standards clear and reasonable. Moreover, it makes clear that the reason why we regulate medical treatments is mostly political, rather than epistemic. We want to minimize the likelihood of a medical catastrophe to happen. As it happened with pharmaceutical regulation, a scandal (Dalkon Shield) makes necessary a political control over the devices, but intuitively testing low-risk ones like rubber gloves in a clinical trial does not sound a very good idea. The political assessment of potential

risks associated with devices dictates the choice of mandatory testing standards of safety and efficacy to grant market approval. This political assessment of risk is purely informal, but it seems enough to justify the existence of multiple standards of testing. However, this political consensus on the acceptable levels of risk depends on a particular conception of it. As we are going to see in the next paragraph, the risk regulators care about can be defined as the likelihood of a treatment of producing a medical catastrophe.

4.5 Classifying risks: hazards and exposure

It seems as if the existence of a multiple testing standard for medical treatments has been socially admissible well before the 21CCA. The admissibility of these various standards depends on the perceived risks involved in each treatment. Politically speaking, this is all our societies demand from our regulatory bodies: stricter testing regimes are, of course, conceivable and defensible (Stegenga 2016; Osimani 2013), but the forty years of regulatory experience between 1962 and 2017 suggest that there is nothing wrong *per se* with a multiplicity of testing standards. Key to this approach is, of course, the correct assessment of the risk involved, and the history of medical device regulation shows that these assessments sometimes fail. In order to defend a risk-based multiplicity of standards, we need to justify the way the risk assessment is carried out.

Over the last years, indeed, some high profile regulatory failures have raised concerns about the evaluation of medical devices (Ciani, Tarricone, and Taylor 2016; Ciani, Federici, and Pecchia 2018; Hines et al. 2010; Zuckerman, Brown, and Nissen 2011). For instance, in 2012, thousands of silicone-gel based breast have been withdrawn from the market following many cases of leaks of silicone inside patients' body (Horton 2012). In 2010, an artificial hip implant (ASR XL Acetabular System) was

recalled after it became clear that "the device failed at the astonishing rate of at least one in eight […]. 21% of these hips have had to be replaced by 4 years after implantation, and the revision rate rises to 49% at 6 years, as compared with 12 to 15% at 5 years for other devices" (Curfman and Redberg 2011). In both cases the devices, considered moderate-risk, entered the market after approval under the 510k process.

In order to understand, and justify, the way regulatory risks are assessed, we suggest adopting a common distinction in the field of chemical management between *hazards* and *risks* (van Leeuwen e Vermeire 2007). The term "hazard" refers to the intrinsic properties of a chemical, and its likelihood to do harm. While, the *risk* is the combination of hazard and *exposure* (R=H*E). The risk is what you want to mitigate by changing either or both components, hazard and exposure. In purely mathematical terms: if one of them is zero, the risk is also zero. Then, for instance, with regard to low risk medical devices, they harbor a hazard near to zero, therefore the assessment of the risk after the exposure is close to zero as well. This means that they will hardly produce a catastrophe in terms of harms, which is why they are assigned to a lower testing standard. This distinction would allow us to explain also why the FDA does not regulate surgical treatments even though intuitively they harbor much more hazards than latex gloves or than an ordinary drug for colds. However, surgical treatments are not administered in the same way of drugs. When a new drug is released on the market, it becomes immediately available for thousands of patients, whereas new surgical procedures are administered to individual patients sequentially. As soon as something goes wrong the surgical procedures can be dismissed, thus the procedure can be extremely harmful, but it eventually harms only few patients. Paradoxically, a dangerous surgical procedure might harm more patients in RCTs than in the real world. Therefore, from a public health perspective, surgical procedures are not much more worrisome for regulators than rubber gloves: calling back the equation R=H*E, since

the exposure is very low the risk is as low as well. Whereas, when both the hazards and the exposure are more than zero then it is necessary to assess the risk trough some experimental procedures which can furnish a better estimate of the level of safety of treatments.

The point of introducing the risk vs. hazard distinction is to make explicit the two major factors behind regulatory risk assessment in medicine. Regulatory bodies like the FDA do not care about hazards *per se*, but in combination with exposure. Regulatory agencies do not have the resources to provide a uniform level of protection for patients whenever there is the possibility of hazards. In allocating testing resources, agencies weight hazards by exposure, although not always in a purely quantitative manner[4]. We will not defend this approach from a normative standpoint, for which there would be many different arguments (e.g. in utilitarian terms: the greatest good for the majority of patients) and objections. For the sake of the argument, it is enough to defend its empirical adequacy: the risk vs. hazard distinction allow us to grasp the political consensus about treatment testing. We want stricter tests where the risks are bigger, and we do not care so much about hazards if the exposure is low. From this standpoint, the multiplicity of testing standards is politically defensible: there seems to be a broad agreement in the US (and in those countries which imitate its regulatory system) about calibrating our testing standards according to this particular concept of risk.

Of course, it is open to debate how to articulate and refine the concepts of hazard and exposure. Hazards do not come in a single flavor, instead they exhibit qualitative differences. Death, for instance, is far superior to any other potential adverse events

[4] E.g. sulfanilamide was an antibacterial compound to treat streptococcal infection that, in the late 1930s, was marketed in a toxic solution that caused more than 100 deaths in the United States. The supporters of granting stronger powers to the FDA framed the scandal in terms of the group of most likeable victims: white, virginal kids avoiding any mention of the black, male, and possibly sexually licentious consumers of sulfanilamide (Carpenter 2010).

(e.g inflammation). Most often, hazards cannot be measured on a unidimensional scale and the exercise of value judgment becomes central. But this does not mean that hazards are incommensurable. We have a consensus on what we take as a hazard, and it depends on the outcome associated with that. Take for instance the definition of serious adverse events that the FDA implements for market withdrawal, or black box warning. An adverse event is any undesirable experience associated with the use of a medical product in a patient. The event is considered to be serious and should be reported to the FDA when the patient outcome is among some of the following: death, life-threatening condition, hospitalization, disability or permanent damage, congenital anomaly or birth defect, intervention to prevent permanent impairment.

As to the exposure[5], we are referring to the number of individuals who can potentially enter in contact with a medical product, that is the target population. Since its birth the FDA has regulated medical drugs recognizing the differences between risks associated with the practice of medicine, which are individual, and risks associated with the mass production of drugs, which are public. As it were, the FDA had always taken into account consideration about exposure, leaving almost unregulated the product or practices with low individual risks while regulating medical products that carry with them risks associated with mass production. As to pharmaceuticals, in 1962, we had undifferentiated populations of pharmaceutical consumers, but we have been refining the target decade after decade. For instance, today our understanding of cancer biology is solid enough to define patients according to the genomic profile of the tumor, rather than its site of occurrence. Using genome sequencing in clinical setting, we can identify previously occult biomarkers of drug sensitivity that can aid in the identification of patients most likely to respond to targeted anticancer drugs. This

[5] Here we are using the term exposure in a more intuitive rather than technical sense. In the medical field it is defined as the amount of a factor (a variable of interests) to which a group or individual was exposed, and it usually captures the temporal aspect (see Velentgas et al. 2013).

makes possible to have target population of hundreds of patients, or even less. Therefore, for this class of drugs the likelihood of a medical scandal is lower than for an ordinary drug, perhaps comparable to medical devices or surgical treatments[6].

As Iyer and collegues demonstrated in their landmark study on everolimus (Iyer et al. 2012), some drugs that might fail in RCTs are actually effective in cancer patients harboring a specific – but rare – somatic mutation. Now the 21CCA opens up the possibility of drug approval for genetically defined groups. For instance, it allows the use of Real World Evidence (RWE) to support a new drug application or a label change. Thus, as of today there is not yet a clear consensus on how real-world evidence should be used, and how it can impact on the approval process. According to some commentators (Sherman et al. 2016), intuitively, collection of RWE is impossible before the approval of a drug. By contrast, we think that collection of RWE is actually feasible. For example, it could be accomplished allowing access to investigational drugs to all the patients who may benefit from them but who do not meet the entry criteria in the clinical trials. This access may be launched for therapies targeting small populations of patients in parallel with phase II trials, hence right after the assessment of the minimum requirements of safety. This program would help investigators to address one major issue in the drug approval process. Rather than running expensive, time consuming, and challenging RCTs, collection of RWE can provide more and faster information about the safety and efficacy of new therapies.

[6] From an epistemic point of view, this approach rehabilitates the role of mechanisms in assessing drug safety and efficacy, which has been instead minimized by many EBM scholars, yet strenuously defended by as many philosophers of science (Clarke et al. 2013). Indeed, the a priori risk-based classification heavily depends on the mechanistic knowledge, both for assessing the hazards and for defining the target population. It is not that the knowledge of mechanisms of action of a medical treatment is necessary to grant an approval, or to easing the testing standards, but it is necessary to assess the likelihood of a drug of producing a medical scandal and therefore to driving the choice of the most appropriate regulatory testing standards.

From a political standpoint, multiple standards of drug testing are acceptable to the extent that we have an agreement on the levels of tolerable risks, and we have indeed the basis of such agreements in our current regulation. Our conceptualization of risks makes more explicit the basis of this consensus. It seems reasonable to associate different testing standards to medical products according to their likelihood of producing a medical scandal. We should not let fears based on pharmaceutical catastrophes that happened more than 50 years ago stop us from improving our drug regulatory system.

4.6 The costs of uncertainty

Of course, there is always a degree of uncertainty around every medical treatment. With regards to this, both surgical interventions and medical devices makes no exception, let alone pharmaceutical treatments. Therefore, one potential objection to the risk-based choice of testing standards is that we should always run the most stringent test in order to reduce our uncertainty around the real risks and benefits of any medical treatment. From this perspective, the adoption of different evidentiary standard for surgery, medical devices, or drug, is not justifiable. This idea however is entrenched in the more general view – epistemically defensible - whereby uncertainty is unacceptable when coming to risks and medical regulatory decisions, and RCTs are the best and the only way to reduce it.

Let us then focus on uncertainty. First of all, it is a very different concept from risk. With regard to regulatory decisions, sometimes we have a tendency to think that they are linked, the more the uncertainty the more the risks for patients, but it is not the case. In fact, a drug can be safe and effective even though we have no clue about it. It might sound trivial, but the risks are independent from the results of any experiment. Second, as well as we rarely have a complete knowledge of mechanisms, we hardly

have a complete statistical knowledge. For instance, it is largely acknowledged that efficacy and safety data obtained in "artificial" setting, such as RCTs, are not necessarily a good predictor of the effects on the real-world environment of clinical use (Cartwright 2010). Therefore, given the gaps in our knowledge, all regulatory decisions regarding whether to license a drug or not are taken under a condition of uncertainty.

The level of acceptable uncertainty around benefits and risks is debatable (Eichler et al. 2013; Moore and Furberg 2012). What is sometimes overlooked in the current epistemological debate about evidentiary standards in drug regulation is the unfortunate fact that in a resources-constrained environment, every piece of additional information comes at an *opportunity cost* (Beckman, Clark, and Chen 2011). The concept of "opportunity costs" can be easier if illustrated by means of an example. Consider the detection of a clinically relevant, but very rare, adverse effect of a drug, such as carcinogenicity. Being the adverse effect very rare, to tell apart if it is a relevant signal or a by-chance product, it is necessary to run clinical trials that would involve tens of thousands of patients. Therefore, this trial would surely reduce our uncertainty about the drug safety, but it would require a large investment for a result that would add little to the regulatory decision. That is, because the risk is so small the benefits may still outweigh this risk. Very often, increasing amounts of investment or effort in clinical research produce very little small gains in knowledge. A more concrete example of the potential opportunity costs of drug regulation is provided by Bouvy and colleagues (Bouvy et al. 2012). In this study, the authors assess "the cost-effectiveness of the International Conference on Harmonisation (ICH) E14 guideline that requires a thorough QT/QTc (TQT) study for all drugs under development". Prolongation of the QT interval in the surface electrocardiogram (ECG), which could result in potentially fatal ventricular tachyarrhythmias, was a leading cause for drug withdrawals during 1988–2000 (Shah 2006). This is why in 2005, the International Conference on

Harmonisation (ICH) promulgated a guideline (ICH E14) calling for a "thorough QT/QTc" (TQT) study for all new drugs before approval. ICH E14 guideline has been adopted by the US Food and Drug Administration, as well by the European Medicines Agency. But this has led to an increase of the costs involved in clinical trials. The authors then compared the costs of regulation compared to non-regulation (that is clinical trials without costs of TQT studied and ECG monitoring). They conclude that ICH E14 costs society €187,000 to gain one QALY (quality-adjusted life year) and €2.4 million to prevent one drug-induced sudden cardiac death. This study highlights the need to determine acceptable levels of risks related to the use of drugs and acceptable cost-effectiveness of safety-related regulatory actions.

The bottom line is that when we discuss the level of uncertainty, we should be aware of the costs involved in any attempt to reduce it, therefore in the choice and development of evidentiary standard. It has been always pointed out – with good reason – that drug regulatory decisions must be separated from the economic considerations. However, given the recent controversy and concerns about the prices of new drugs, many are starting to realize that regulatory authorities cannot ignore anymore the cost implications of their licensing requirements, because those costs will ultimately be passed on to consumers. As Bouvy and his coauthors put it, "in a world of rising health-care expenditures and increasing drug development costs, regulatory agencies and society at large should think carefully about what they are willing to pay for reassurance with respect to drug safety. This is particularly relevant if determining these risks does not ultimately translate into substantial health gains, or when health gains can be achieved only by spending vast amounts of money" (Bouvy et al. 2012).

In addition, we should also consider non-financial costs, which instead impinge exclusively on patients: the costs of withholding a drug. Recently a pair of MIT economists argued that the one-size-fits-all approach of the FDA to the approval

process for drugs (from immunotherapy to flu) is a non-sense. As a result, the FDA is too conservative in regulating drugs for severe disease like lung cancer and too lax for less dire one like prostate cancer. Andrew Lo and his MIT colleagues (Montazerhodjat et al. 2017) focused on the statistical rationale of RCTs, and in particular on the level of type I error (i.e. the threshold of false positive results) that the FDA tolerate, which as everyone knows is fixed at 2.5%. The authors think that this threshold should instead vary from disease to disease, according to both its severity and its prevalence. This is based on the idea that patients who suffer from more serious conditions would be more willing to accept more uncertainty, because the alternative is often the death. While someone with a less severe disease such as diabetes presumably cares more about avoiding potential side effects, and therefore she would like to be more certain about the risks involved in new treatments. Thus, they perform a Bayesian decision analysis to estimate the threshold of type I error the FDA should accept for different drugs. Basically, even though they do not take into account the *hazards* of drugs, and they do not consider evidentiary standards alternative to RCTs, they supply a concrete example of how a risk-based classification of testing standard should be organized. For example, since pancreatic cancer is one of the most severe disease, they proposed a false positive rate for it of 28%. Although, according to their model the size of population is a positive modifier of the severity, and it actually increases the type I error threshold, whereas we would argue the opposite.

Anyway, their analysis shows how the will to accept the uncertainty in testing medical treatments can be taken into account in choosing the most appropriate method. However, even if intuitively it might sound appealing, there is limited information about the levels of uncertainty that patients are willing to accept or tolerate, and we do not know whether and how that would differ from the one of regulators. With regard to this, there are some case studies, which are exclusively

anecdotal though. These cases are usually about drugs for untreated diseases that, once released, show some expected serious adverse event. The FDA therefore withdraws the drug from the market, and patients start to complain about that decision. For instance, natalizumab, a drug for multiple sclerosis, was approved in late 2004. Soon after, some physicians reported cases of progressive multifocal leukoencephalopathy (PML), a life-threating condition, in patients receiving the new drug, and the FDA decided to withdraw it despite the great uncertainty around both the incidence and the genuine causal link between the drug and the adverse event. However, in 2006 the drug was reintroduced to the market due to the pressing request of patients and family members. A survey of patients with multiple sclerosis (Calfee 2006) found that the majority would get a drug which might be more effective than the currently available drugs, even if it had a chance of one in a thousand of causing fatal side effects. In another survey of patients taking natalizumab (Miller, Karpinski, and Jezewski 2012), patients complained that the risks of natalizumab were overemphasized and that its potential benefits were overlooked. This case also shows how difficult it is to reduce the uncertainty about the adverse drug reactions, especially if they are rare. For instance, after many years, the incidence of PML for natalizumab is estimated to be >1/1000 to <1/100, still far to being precise.

4.7 Conclusion

The recent enactment of the 21st Century Cures Act is ushering a new season of heated debates about regulatory evidentiary standards. Most of the conservatives are worried by the adoption of different standards of testing for certain drugs, because of the potential negative impact that it would have on patients. We have shown that double standards are, as a matter of fact, pervasive in medicine, and that we accept them because we are willing to accept the consequences of potential mistakes. We have

focused on the case of surgical treatments, which are no regulated at all. We have argued that the lack of regulatory imposition is due neither to epistemic nor political reasons, since all of them are defeasible. Why then there is no an FDA for surgery? We have tried to offer an answer analyzing instead the medical devices regulation, which allow some devices to get market-approval almost without any proof of safety and efficacy (e.g. rubber gloves), while it demands more rigorous testing for other devices (e.g. cardiac pacemakers). We have argued that the reason for the regulators to consider different testing standards is political, and lies in the fact that some devices are intuitively safer than others and therefore it is less likely that they will produce a medical disaster. In an environment with limited resources, this approach sounds rational.

Then we have explored the question of how to classify the risks of medical treatments, including drugs. With regard to this, we have suggested to adopt a distinction between hazards and risks in order to conceptualize the likelihood for a drug of producing a medical scandal. Taking in account both the hazards and the exposure we can have an a priori estimate of that likelihood. In this context, the size and the definition of target population play a key role. With the progress of biological science, nowadays we can circumscribe the target population for certain kind of drugs, such as cancer targeted therapies. In some cases, this population is composed by few hundreds of patients, therefore from a political standpoint potential adverse effects do not pose a threat to public health.

Finally, we have considered a potential objection to our risk-based approach to regulatory evidentiary standards: accepting lower standard of evidence paves the way for more regulatory failures. However, this is hard to defend without the lack of an empirical benchmark. Of course, the success of this approach strictly depends on our ability in assessing a priori the risks of a drug. And this ability might be impaired by the

gaps in our knowledge, which render our estimate of risks very uncertain. And yet, testing all the medical treatments with RCTs is not the more rational approach to manage this uncertainty. Finally, we provided a concrete example of how this can be done, taking into account eventually also the willingness of the patient to accept different degrees of uncertainty.

In conclusion, there is room for adopting more flexible testing standards in some cases without losing much safety for population. The higher the likelihood of a medical product to generate a scandal the higher the testing standards should be. As suggested by a recent Nature's editorial[7], regulators must collaborate closer with scientists in order to grasp how the new pharmaceutical treatments actually works, and which are the potential hazards that they carry and the patients which could benefit from those treatments. In this way, we could find the best regulatory balance between safety, patient's protection, and innovation, at the best of our scientific knowledge.

References

Avorn, Jerry, e Aaron S. Kesselheim. 2015. «The 21st Century Cures Act — Will It Take Us Back in Time?» *New England Journal of Medicine* 372 (26): 2473–75. doi:10.1056/NEJMp1506964.

Beckman, Robert A., Jason Clark, e Cong Chen. 2011. «Integrating predictive biomarkers and classifiers into oncology clinical development programmes». *Nature Reviews Drug Discovery* 10 (10): 735–48. doi:10.1038/nrd3550.

Bouvy, J C, M A Koopmanschap, R R Shah, e Huub Schellekens. 2012. «The cost-effectiveness of drug regulation: the example of thorough QT/QTc studies». *Clinical Pharmacology & Therapeutics* 91 (2). Wiley Online Library: 281–88.

[7] https://www.nature.com/articles/d41586-017-02231-z

Calfee, John E. 2006. *A representative survey of MS patients on attitudes toward the benefits and risks of drug therapy*. AEI-Brookings Joint Center for Regulatory Studies.

Carpenter, Daniel. 2010. *Reputation and power. Organizational image and pharmaceutical regulation at the FDA*. *Journal of Chemical Information and Modeling*. Vol. 53. Princeton: Princeton University Press. doi:10.1017/CBO9781107415324.004.

Cartwright, Nancy. 2010. «What are randomised controlled trials good for?» *Philosophical studies* 147 (1): 59–70. doi:10.1007/s11098-009-9450-2.

Ciani, Oriana, Carlo Federici, e Leandro Pecchia. 2018. «The evaluation of medical devices: are we getting closer to solve the puzzle? A review of recent trends». In , 916–19. doi:10.1007/978-981-10-5122-7_229.

Ciani, Oriana, Rosanna Tarricone, e Rod S Taylor. 2016. «Comparing Drug and Nondrug Technologies in Comparative Effectiveness Research». *Comparative Effectiveness Research in Health Services*. Springer, 275–90.

Clarke, Brendan, Donald Gillies, Phyllis Illari, Federica Russo, e Jon Williamson. 2013. «The evidence that evidence-based medicine omits». *Preventive Medicine* 57 (6): 745–47. doi:10.1016/j.ypmed.2012.10.020.

Curfman, Gregory D., e Rita F. Redberg. 2011. «Medical Devices — Balancing Regulation and Innovation». *New England Journal of Medicine* 365 (11): 975–77. doi:10.1056/NEJMp1109094.

Deyo, Richard A. 2004. «Gaps, tensions, and conflicts in the FDA approval process: implications for clinical practice». *The Journal of the American Board of Family Practice* 17 (2). Am Board Family Med: 142–49.

Eichler, Hans-Georg, Brigitte Bloechl-Daum, Daniel Brasseur, Alasdair Breckenridge, Hubert Leufkens, June Raine, Tomas Salmonson, Christian K. Schneider, e Guido Rasi. 2013. «The risks of risk aversion in drug regulation». *Nature Reviews Drug Discovery* 12 (12): 907–16. doi:10.1038/nrd4129.

Farrokhyar, Forough, Paul J Karanicolas, Achilleas Thoma, Marko Simunovic, Mohit Bhandari, P J Devereaux, Mehran Anvari, Anthony Adili, e Gordon Guyatt. 2010. «Randomized controlled trials of surgical interventions.» *Annals of surgery* 251 (3): 409–16. doi:10.1097/SLA.0b013e3181cf863d.

Frader, JOEL E, e D A Caniano. 1998. «Research and innovation in surgery». In *Surgical ethics*, 216–41. Oxford University Press, New York.

Freeman, Michael, e Mitchell Fuerst. 2012. «Does the FDA have regulatory authority over adult autologous stem cell therapies? 21 CFR 1271 and the emperor's new clothes». *Journal of Translational Medicine* 10 (1): 60. doi:10.1186/1479-5876-10-60.

Hines, Jonas Zajac, Peter Lurie, Eunice Yu, e Sidney Wolfe. 2010. «Left to their own devices: breakdowns in United States medical device premarket review». *PLoS medicine* 7 (7). Public Library of Science: e1000280.

Horton, Richard. 2012. «Offline: A serious regulatory failure, with urgent implications». *The Lancet* 379 (9811). Elsevier: 106.

Issa, Amalia, Kathryn Phillips, Stephanie Van Bebber, Hima Nidamarthy, Karen Lasser, Jennifer Haas, Brian Alldredge, Robert Wachter, e David Bates. 2007. «Drug Withdrawals in the United States: A Systematic Review of the Evidence and Analysis of Trends». *Current Drug Safety* 2 (3): 177–85. doi:10.2174/157488607781668855.

Iyer, Gopa, Aphrothiti J Hanrahan, Matthew I Milowsky, Hikmat Al-Ahmadie, Sasinya N Scott, Manickam Janakiraman, Mono Pirun, et al. 2012. «Genome sequencing identifies a basis for everolimus sensitivity.» *Science (New York, N.Y.)* 338 (6104): 221. doi:10.1126/science.1226344.

Kesselheim, Aaron S., e Jerry Avorn. 2017. «New "21st Century Cures" Legislation». *JAMA* 317 (6): 581. doi:10.1001/jama.2016.20640.

Leeuwen, Cornelis J van, e Theodorus Gabriel Vermeire. 2007. *Risk assessment of chemicals: an introduction*. Springer Science & Business Media.

Lilford, R., D. Braunholtz, J. Harris, e T. Gill. 2004. «Trials in surgery». *British Journal of Surgery* 91 (1): 6–16. doi:10.1002/bjs.4418.

McCulloch, P., I. Taylor, M. Sasako, B. Lovett, e D. Griffin. 2002. «Randomised trials in surgery: Problems and possible solutions». *British Medical Journal* 324 (7351).

Miller, Colleen E, Mary Karpinski, e Mary Ann Jezewski. 2012. «Relapsing-remitting multiple sclerosis patients' experience with natalizumab: a phenomenological investigation». *International journal of MS care* 14 (1). The Consortium of Multiple Sclerosis Centers: 39–44.

Montazerhodjat, Vahid, Shomesh E. Chaudhuri, Daniel J. Sargent, e Andrew W. Lo. 2017. «Use of Bayesian Decision Analysis to Minimize Harm in Patient-Centered Randomized Clinical Trials in Oncology». *JAMA Oncology* 3 (9): e170123. doi:10.1001/jamaoncol.2017.0123.

Moore, Thomas J., and Curt D. Furberg. 2012. «The Safety Risks of Innovation». *JAMA* 308 (9): 869. doi:10.1001/jama.2012.9658.

Osimani, Barbara. 2013. «The precautionary principle in the pharmaceutical domain: a philosophical enquiry into probabilistic reasoning and risk aversion». *Health, Risk & Society* 15 (2): 123–43. doi:10.1080/13698575.2013.771736.

Petrini, Carlo. 2013. «Surgical experimentation and clinical trials: differences and related ethical problems». *Annali dell'Istituto Superiore di Sanità* 49 (2). SciELO Public Health: 230–33.

Sacks, Chana A., Jerry Avorn, e Aaron S. Kesselheim. 2017. «The Failure of Solanezumab — How the FDA Saved Taxpayers Billions». *New England Journal of Medicine* 376 (18): 1706–8. doi:10.1056/NEJMp1701047.

Shah, Rashmi R. 2006. «Can pharmacogenetics help rescue drugs withdrawn from the market?» *Pharmacogenomics* 7 (6): 889–908. doi:10.2217/14622416.7.6.889.

Sherman, Rachel E., Steven A. Anderson, Gerald J. Dal Pan, Gerry W. Gray, Thomas Gross, Nina L. Hunter, Lisa LaVange, et al. 2016. «Real-World Evidence — What Is It and What Can It Tell Us?» *New England Journal of Medicine* 375 (23): 2293–97. doi:10.1056/NEJMsb1609216.

Solomon, Michael J., e Robin S. McLeod. 1995. «Should we be performing more randomized controlled trials evaluating surgical operations?» *Surgery* 118 (3): 459–67. doi:10.1016/S0039-6060(05)80359-9.

Stegenga, Jacob. 2016. «Measuring Harms». In *The Routledge Companion to Philosophy of Medicine*. Routledge. doi:10.4324/9781315720739.ch31.

The EC/IC Bypass Study. 1985. «Failure of Extracranial–Intracranial Arterial Bypass to Reduce the Risk of Ischemic Stroke». *New England Journal of Medicine* 313 (19): 1191–1200. doi:10.1056/NEJM198511073131904.

Tobbell, Dominique. 2011. *Pills, power, and policy: the struggle for drug reform in cold war America and its consequences*. Vol. 23. Univ of California Press.

Tonelli, Mark R., Joshua O. Benditt, e Richard K. Albert. 1996. «Clinical experimentation: Lessons from lung volume reduction surgery». *Chest*. doi:10.1378/chest.110.1.230.

Velentgas, Priscilla, Nancy A Dreyer, Parivash Nourjah, Scott R Smith, e Marion M Torchia. 2013. *Developing a protocol for observational comparative effectiveness research: a user's guide*. Government Printing Office.

Zuckerman, Diana M, Paul Brown, and Steven E. Nissen. 2011. «Medical Device Recalls and the FDA Approval Process». *Archives of Internal Medicine* 171 (11). doi:10.1001/archinternmed.2011.30.

Zuckerman, Diana M, Nicholas J Jury, and Christina E Silcox. 2015. «21st Century Cures Act and similar policy efforts: at what cost?»: *BMJ*, novembre, h6122. doi:10.1136/bmj.h6122.

5. STATISTICAL EVIDENCE AND THE RELIABILITY OF MEDICAL RESEARCH

"Everything ritualistic must be strictly avoided, because it immediately turns rotten" (Ludwig Wittgenstein, Culture and Value)

5.1 Introduction

Statistical evidence is pervasive in medicine. In this chapter, we will focus on the reliability of randomized controlled trials (RCTs) conducted to test the safety and efficacy of medical treatments. RCTs are scientific experiments and, as such, we expect them to be replicable: assuming that the result is true, if we repeat the same experiment time and again, we should obtain the same outcome (Norton 2015). The statistical design of the test should guarantee that the observed outcome is not a random event, but rather a real effect of the treatments administered. However, for more than a decade now we have been discussing a replicability crisis across different experimental disciplines including medicine: the outcomes of trials published in very prestigious journals often disappear when the experiment is repeated - see for instance (Lehrer 2010; Begley and Ellis 2012; Horton 2015).

There are different accounts of the reason for this replicability crisis, ranging from scientific fraud to lack of institutional incentives to double-check someone else's results. In this chapter we will use the replicability crisis as a thread to introduce some central issues in the design of scientific experiments in medicine. First, in section 1 we will see how replicability and statistical significance are connected: we can only make sense of the *p*-value of a trial outcome within a series of replications of the test. But in

order to conduct these replications properly, we need to agree on the proper design of the experiment we are going to repeat. In particular, we need to prevent the preferences of the experimenters from biasing the outcome of the experiment. If there is such a bias, when the experiment is replicated by a third party, the observed outcome will vanish. In section 2, we will argue that trialists need to agree on the debiasing procedures and the statistical quality controls that feature in the trial protocol, if they want the outcome to be replicable. In section 3 we will make two complementary points. On the one hand, replicability *per se* is not everything: we need trial outcomes that are not only statistically significant, but also clinically relevant. On the other hand, trials are not everything: the experts analyzing the evidence can improve the reliability of statistical evidence, although they sometimes fail; we need to study further how they make their decisions. In section 4 we will use a controversy about the over-prescription of statins to show how non-replicable effects are obtained in trials and how experts may fail at detecting such flaws, if the commercial interests at stake are big enough.

5.2 What sort of statistical evidence is the *p*-value of a trial?

Mathematical statistics, with different degrees of sophistication, has been used for different purposes in medicine since the 19th century (Matthews 1995). One major purpose has been the assessment of the efficacy of treatments and a significant step forward in our ability to assess this efficacy was the implementation of the RCT as a testing standard in the 1940s (Marks 1997). The RCT is an experimental design articulated by the statistician Ronald A. Fisher in the 1930s, endowing a comparative method for causal inference with statistical foundations that allowed an interpretation of the outcome (Armitage 2003). In its simplest form, an RCT assesses the effect of a treatment on a given population comparing it to a standard alternative or a placebo[1].

[1] See (Hackshaw 2009) for a quick overview.

The treatments are randomly allocated to the individuals in the test, usually an equal number in each treatment group. After the administration is complete, we measure the variable of interest to assess whether there is any significant difference between the two groups of patients.

In order to quantify the significance of the observed difference, Fisher arranged the experiment as a test of the hypothesis that there is no difference between the two treatments (Teira 2011). This latter is known as the "null hypothesis". Under this assumption, you can calculate the probability distribution of all potential outcomes of the experiment. In other words, a statistically significant difference is an outcome for which the probability, under the null hypothesis, is very low. Fisher introduced as an index of significance the so-called *p*-value, the probability of obtaining a result as extreme as the observed trial outcome or more if there is indeed no difference between treatments. A *p*-value of 0.05 means that, assuming that the null hypothesis is true, if you repeat the trial time and again, only in 5% of the repetitions will you observe such an extreme outcome or an even more extreme one.

If you obtain a statistically significant result, with a *p*-value below the conventional threshold of 0.05, there are two possible ways to interpret this outcome: either the initial hypothesis is true (there is no difference between treatments) and you have observed a rare event, or, the event is actually not rare at all and the hypothesis is just false. There is no way to tell which is the case other than replicating the experiment and seeing whether further outcomes confirm or disconfirm the hypothesis that there is no difference between the effects of both treatments. If repeated trials of the experiment continue to give "unexpected" results, the therapy probably works and the null hypothesis is probably false. If most trials give no significant difference, then the trial that did so was probably just a fluke, and the null hypothesis is probably true. Thus, ultimately, drawing conclusions from clinical trials

is an inductive inference: you are trying to prove the truth of a general proposition (or its negation) on the basis of a finite number of instantiations. There is no surefire method to decide whether the hypothesis is actually true or not. As Ronald Fisher put it, one has a real phenomenon when one knows how to conduct an experiment that will rarely fail to give a statistically significant result: we can show time and again that there is a real difference between the effects of the tested treatments (Spanos and Mayo 2015).

We should notice a crucial point in this argument. The *p*-value estimates how often an outcome will appear in a series of replications of the experiment. Thus, Fisher's interpretation of the trial outcome requires a *frequentist* understanding of probabilities as opposed to a Bayesian approach where probabilities measure our degree of belief in the truth of a given statement[2]. A Bayesian trial would measure how strong our belief in the safety and efficacy of a treatment is. In a frequentist trial we measure instead how often we will observe a given outcome if we repeat the same experiment time and again. Our *p*-values are tied to an experimental design. If we conduct a somewhat different trial of the same therapy, the probability distributions of outcomes will be different, and thus an outcome that was statistically surprising in the original experiment may not be in the new one. Thus, paradoxically, identical outcomes in two differently designed experiments may not confirm each other. Confidence intervals, alpha values and other frequentist concepts for hypothesis testing are equally tied to an experimental plan.

As a general epistemic principle, scientific experiments should be replicable: if we implement the same design properly, we should obtain the same outcome independently of any subjective feature of the experimenter or the contingent

[2] See (Nardini 2016) for a quick overview.

circumstances of the experimental setup. The more replicable an outcome, the more reliable it is. In clinical trials, as in other fields in science, *p*-values provide an implicit index of the replicability of an outcome: if we reject a hypothesis about both treatment effects being equal, we should expect the new treatment to perform better than the alternative whenever we administer it according to the trial protocol (patients, dosage, etc.). However, as we are will see in the next section, the *p*-value may be a misleading index of replicability.

5.3 The sources of non-replicability

In 1962, the US Food and Drug Administration (FDA) received the mandate to test the safety and efficacy of new treatments with "well controlled investigations," later specified as two RCTs plus one further confirmatory trial (Carpenter 2010). This new regulatory standard created the contemporary trial industry, with pharmaceutical companies heavily investing in the design and conduct of RCTs in order to gain market access for their compounds. The FDA experts are supposed to assess these trials and infer whether the outcome observed in the sample of patients participating in the trials will obtain when the treatment is used on the general population. In other words, the FDA experts should assess the *external validity* of the trial (see La Caze 2016) that is, whether the causal connection established in the trial between the treatment, on the one hand, and improved patient outcomes, on the other, will hold in non-experimental clinical settings. If the drug is approved and then turns out not to be safe and efficacious – e.g. if unexpected adverse effects are observed once the treatment is released commercially – we would have accepted the wrong hypothesis in the trial: the experimental treatment would actually be inferior to the standard alternative.

A correct decision should be grounded on reliable trial outcomes and in order to obtain these latter, the experimenters testing a drug should agree, at least, on the

proper controls to be implemented in the trial and on the adequate statistical design of the experiment. Otherwise, the *p*-values of their trials may be pointing to different experimental designs, providing non-comparable evidence. Ideally, a good trial should be *internally valid* (La Caze 2016): the experimental protocol should properly capture the causal connection between the administered treatment and the observed effect. A correct causal inference should be grounded in a *like with like* comparison. The different arms of the trial should be entirely alike except for the treatment each group of patients receives. Othewise, we would be unable to tell whether the observed difference between treatments originates in the causal effect of each treatment or in a non-controlled factor that creates a difference between groups. For several centuries, physicians have been debating the proper experimental *controls* that a *fair* test should implement in order to fend off confounding factors. The reader should bear in mind that this is an endless debate (Franklin 1990): every experimental setup is different and so are the potential confounding factors and the corresponding controls. Experimenters in all disciplines have their own checklists updated according to the progress in their fields.

In medicine, researchers have paid particular attention to the biases originating in the preferences of either the experimenters or the experimental subjects and how to control for them. Non-replicable outcomes are usually blamed on these sort of biases: the interests of the pharmaceutical industry spoils the design of their sponsored trials, so that their outcome disappears once these tests are conducted in an unbiased manner. There are a large number of biases (Bero and Rennie 1996) so we can only illustrate here some that are particularly relevant for the replicability crisis. We will focus on two stages of the experiment: the conduct of the test and its statistical interpretation.

As to the former, there is a clear consensus on some of the biases that may spoil a trial outcome. Selection bias occurs when the allocation of subjects to study groups is contaminated by the preferences of the experimenter (e.g. the healthiest patients receive the experimental treatment). Usually it occurs when recruiters selectively enrol patients into the trial based on what the next treatment allocation is likely to be (Kahan, Rehal, and Cro 2015). Randomization controls for selection bias and is therefore considered a pre-requisite for a methodologically sound trial. So is the masking of treatments, so that the physicians and patients in the trial cannot ascertain what they are giving or getting, guaranteeing that their preferences do not bias the treatment effect. However, there is still no consensus on the full list of controls that should be implemented in a trial in order to consider it unbiased.

Peter Gøtzsche (Gøtzsche 2013) illustrates the risk of unmasked trials as follows. In trials of antidepressant drugs, we usually assess subjective outcomes, even if the assessor is often a third party and not the patient himself. There is evidence from a meta-analysis (Hróbjartsson et al. 2014), that when the assessor is not masked to the treatment patients receive (i.e., she knows whether they got the experimental drug or a placebo), the assessor overestimates the effect on average by 36%.

Reaching statistical significance is often a matter of getting a few more positive outcomes. Following Gøtzsche (Gøtzsche 2013), if you are testing an antidepressant versus a placebo on 400 patients, the *p*-value of observing 19 more patients improve with the experimental drug than with the control is 0.07

	Improved	Not improved	Total
Drug	119	81	200
Placebo	100	100	200

However, if you observe two more patients improve with the active treatment (121 instead of 119), then your trial will reach statistical significance ($p = 0.04$). A non-

masked assessment of outcomes increases thus the chances of getting a positive result. We may suspect that failing to mask the assessor could have been intentional if the sponsor of the trial was seeking such a favorable outcome. Here we see what is at stake with the *internal validity* of the trial: the design of the experiment may fail to grasp the causal connection (or rather, in this case, lack thereof) between the treatment and its study's outcome, with the *p*-value providing misleading evidence about the treatment efficacy.

Biases, which by their nature do not (necessarily) repeat each time a trial is redone, can thus be a cause of non-replicability. If we wish to eliminate bias, we need to agree on the list of proper controls that would guarantee an unbiased outcome and incorporate them into the trial protocols, in order to maximize our chances to observe the same outcome whenever we repeat the experiment. How far are we from these ideal list of debiasing controls? In principle, we should aim at controlling for every source of human intervention, but this is difficult to achieve. For instance, Claes-Fredrik Helgesson (Helgesson 2010) has illustrated practices of out-of-protocol data cleaning in large Swedish RCTs. Helgesson tracks the ways in which data are informally recorded and corrected without leaving a trace in the trial's logbook, from post-it notes to guesses about the misspelling of an entry. In his view, those who make such corrections do so in good faith, in order to increase the credibility of their results. Would these corrections threaten the internal validity of the outcome? After all, if the experiment was replicated elsewhere, the corrections might be different and the test would yield a different outcome. But if we tried to explicitly control for these cleaning practices the experimental protocol would become extremely cumbersome. This is why it is so difficult to agree on a full list of controls: experimenters have different standards as to what constitutes an unbiased experiment and we need to reach a

compromise in between absolutely unbiased (but unfeasible) protocol and protocols that are too open to interested manipulations.

As we noted above, the statistical analysis of trial results, as well as the study design itself, can lead to problems in replicability, as statististical analyses can also be biased, (e.g. according to the preferences of the sponsor) most notoriously when the sample size is not chosen according to statistically justified principles. In biomedical research, a particularly vocal critic of this statistical flaw is the epidemiologist John Ioannidis. Although some of his claims are controversial, such as "most published research findings are false" (see Sorić 1989; Ioannidis 2005; Ioannidis 2006; Ioannidis 2008; Ioannidis 2014), his contributions are worth considering as a focal point in the replicability debate. Take for instance his empirical evaluation of very large treatment effects (VLE) of medical interventions (Pereira, Horwitz, and Ioannidis 2012). A standard complaint about industry sponsored research is that trials are designed to detect small treatment effects that would guarantee regulatory approval without any clinical innovation (e.g. "me too" drugs): in principle, VLE would sort out this problem. Ioannidis and his coauthors define a statistical threshold for VLE, and used data from the Cochrane Database of Systematic Reviews to identify studies that showed such effects and track further studies on such outstanding outcomes. They found that VLEs usualy arise in small trials with few events, and their results typically become smaller or even lose their statistical significance as additional evidence is obtained. According to Ioannidis (Ioannidis 2008), we have here a problem of statistical literacy: biomedical researchers tend to claim discoveries based exclusively on *p*-values, focussing on significance while ignoring statistical power, which is a measure of whether a study is large enough to detect what it is looking for. Without a proper sample size, it is impossible to tell a random spike in the data from a true treatment effect. If the sample is small, we may observe a large difference by chance, but if the experiment were

repeated and the sample size grew, chance would gradually give way to the true treatment effect – see for instance (Button et al. 2013). Replicability fails to obtain because there might have been no effect to grasp – even if the trial protocol itself was unbiased. Although adequate sample size is usually included in lists of requirements for well designed studies, it is still often not met, as not all medical journals require it for publication. As before, part of the problem is lack of agreement as to which tools for bias control to require of researchers.

Summing up, biases can contaminate the trial and spoil the statistical reliability of the outcome both while the experiment is being conducted and when the data are interpreted. The replicability of a trial will depend on which debiasing procedures and statistical quality controls that experimenters adopt in their experimental protocols. The more replicable the trial, the more reliable the information it yields.

5.4 Is the problem truly a crisis?

Although we have discussed some of the sources of the replicability crisis, the question remains whether it is reasonable to refer to the problems we have with replicability as a crisis. On the one hand, a trial may be replicable and yet it may not deliver the information we actually need: we want clinical, not just statistical reliability. Replicability is no guarantee of clinical benefit. On the other hand, despite the problem with the replicability of trials, regulators seem to have coped with them reasonably well until recently, according to the available data. In other words, even without replicability, expert judgment has allowed us to make proper decisions about the safety and efficacy of drugs.

Let us argue for the first point: (Pereira, Horwitz, and Ioannidis 2012) note that VLE usually appear with treatments whose efficacy is defined by a laboratory test (e.g. hematologic response), as opposed to a clinically-defined efficacy (e.g. symptomatic

improvement) or a fatal outcome (e.g. death). There were only three reliably documented VLE that used mortality as an endpoint (out of 2791). We see here another contentious point in contemporary debates on biomedical research: sometimes there are good reasons to adopt *soft* endpoints (e.g. biological or imaging biomarkers) instead of *hard* trial outcomes (mortality data); sometimes not (Asmar and Hosseini 2009). According to the industry critics, *soft* endpoints are chosen in order to get a statistically significant effect of a treatment, even if it is clinically not very interesting. This positive effect is just enough for the manufacturing company to request regulatory approval. Such trials may be unbiased, statistically well-grouned and perfectly replicable, but the research question they are addressing may just concern the commercial interest of the manufacturer sponsoring the trial rather than the clinical interests of patients and physicians alike –as we will see in our case study below. This point suggests that some of the issues at stake in the replicability crisis go beyond the methodological quality of trials as scientific experiments and rather pertain to their clinical goals: what trial outcomes should we look for and who should decide about them?

Let us argue for our second point now: expert judgment can improve the reliability of the information provided by trials. If trials were systematically unreliable, the decisions of regulatory agencies such as the FDA would be systematically misguided. Critics like Gøtzsche (Gøtzsche 2013), for instance, think that this is actually the case: 70% of FDA scientists are not confident that the drugs they approve are safe. If the internal or external validity of a trial fails, we will indeed observe outcomes in the population that were not anticipated in the trial.

Dan Carpenter has tracked such unanticipated outcomes through label changes: adverse effects observed in the commercial use of a drug are often incorporated into its brochure. From 1980 to 2000, the average drug received five labeling revisions,

about one for every three years of marketing after approval (Carpenter 2010). Clearly, there is much about the full range of effects of a drug that we only discover after it reaches the market. Regulatory trials are testing the safety and efficacy of a compound, so these new findings do not necessarily call the original studies and their evaluation into question. Indeed, if we judge the reliability of trials by the number of market withdrawals due to serious adverse effects, the figures seem more promising: between 1993 and 2004, only 4 out of the 211 authorized drugs (1.9%) were withdrawn (Carpenter, Zucker, and Avorn 2008). In other words, the external validity of trials might be far from perfect (they don't track the full range of effects), but when it matters (serious adverse effects), the FDA seems to have been making the right decision. How is this possible?

The FDA combines the statistical evidence of clinical trials with expert deliberation: decisions about drugs are not made on the basis of RCTs alone, but in committees with adversarial confrontation of experts (Urfalino 2012). These committees seem to be able to make correct decisions as to the safety and efficacy of drugs and ponder the reliability of the evidence provided by trials –for a critical discussion, (see Stegenga 2016). At least, under certain conditions: a 1.9% error rate (drug withdrawal) in a decade seems a reasonable standard. But when the FDA committee was given a shorter deadline, still in the same period (1993-2004), 7% of the drugs approved were later withdrawn (Carpenter, Zucker, and Avorn 2008). In other words, under certain conditions, expert judgment can improve the reliability of the information RCTs when it comes to making decisions about medical treatments. Further investigation is needed as to how these expert judgments work, but the effect cannot be discounted.

5.5 Case study: a controversy over statins

Let us illustrate with a case study two of the previous points: not large enough trials and the relevance of expert judgment. The treatment under discussion will be statins, a class of drugs that inhibits the cholesterol synthesis associated with cardiovascular diseases (CVD). Statins have been widely used over the last thirty years to prevent CVD, with excellent success in many different trials – and an equally successful record in sales. However, there is a growing concern that statins are being overprescribed on the basis of trials that verify their ability to decrease cholesterol in many groups of patients without evaluating whether they prevent these patients' death – (see Goldacre 2014; González-Moreno, Saborido, and Teira 2015). The reader should bear in mind that this is a controversial issue and the question is far from settled.

This concern about overpresciption was highlighted by the controversy that followed the publication, in November 2013, of The American College of Cardiology/American Heart Association guidelines on the topic. These new guidelines recommend the use of statins for primary prevention of CVD (prevention of CVD in patients who do not yet have it) in patients with a 10-year predicted risk of CVD of 7.5% or greater; statin therapy was suggested as an option in patients with a predicted risk between 5% and 7.4%. These are very low thresholds and, consequently, more than 45 million (about one in every three) middle-aged asymptomatic Americans qualified for treatment with statins. If we consider that the US population is about one-twentieth of the global population in the same age-range, and assuming that the distribution of risk profiles is similar, this would suggest that approximately one billion people should take statins. In Ioannidis's words, this would amount to a "statinization" of the planet (Ioannidis 2014)

Taking statins is not completely harmless: there are side effects, such as myalgia that paradoxically prevents exercise training, which on its own can results in health

gains and decreased CVD (Macedo et al. 2014). So, what were the grounds for such a massive public health intervention? According to Ioannidis (Ioannidis 2014), the guidelines were based on trials that tracked the cholesterol reduction in patients, but did not follow them for long enough to see whether such reductions lowered also their mortality rate. This was the case of JUPITER, one of the biggest trials testing a statin in patients who had not yet shown evidence of CVD (primary prevention) (Mora and Ridker 2006). It showed that the treatment significantly reduced the risk of myocardial infarction, stroke and vascular events, but, because it showed strong evidence of benefit early, the trial stopped following patients after 1.9 years instead of the planned 4 years, and thus was unable to detect an effect on mortality in the participants (De Lorgeril et al. 2010).

Trials are statistically designed to reveal a treatment effect of a given size with a minimal error rate. We need a certain amount of data (a designated number of patients: the sample size) to minimize error. If we interrupt the trial, we are losing data and we can only be certain of identifying the true effect of a treatment under a number of statistical assumptions. JUPITER was interrupted because the preventive effect of statins was judged big enough to make the remaining two years of data accumulation unnecessary. In other words, the implication was that if someone were to try to replicate JUPITER in full, she would observe the same effect, as the effect JUPITER observed was so large, even before it was completed, that it could not reasonably be supposed to be due to chance.

But, in fact, when other researchers tried to reproduce the same effect, they were unsuccessful. For instance, CORONA trial (Kjekshus et al. 2007) aimed to test the efficacy of statins in secondary prevention, treating patients who already have had a cardiac event, with a view to reducing the probability of a second one. The conclusion was that "there were no significant differences between the two groups in the coronary

outcome or death from cardiovascular cause." This was an unexpected outcome, since the trial population should clearly benefit from the preventive effects of statins. Indeed, the physio-pathological mechanism of stroke or myocardial infarction is always the same, statins should be at least as efficacious in the secondary prevention as in the primary and we have not any scientific reason to think the opposite. In fact, the only difference between the two populations is the probability of observing an infarction, which is obviously higher in patients who already had one than in healthy people. This has an important consequence in designing and performing trials. As we have just mentioned in primary prevention, if the population is at lower risk, this means that the probability of observing a myocardial event is low; therefore, the detection of the outcomes needs both a bigger sample size and a longer follow-up of patients. Whereas in secondary prevention, we need less people and a shorter follow-up to show an effect of statins since the probability of observing a cardiac event is high. Therefore, from a statistical point of view, it should be easier to demonstrate the efficacy of statins in secondary prevention than in primary, yet this did not happen. The negative results of CORONA were also reached by two more trials: GISSI-HF (GISSI-HF investigators, 2008) and AURORA (Fellström et al. 2009). In patients undergoing hemodialysis with high cardiovascular risk, rosuvastatin lowered the LDL cholesterol level but had no significant effect on a *hard* composite end point (death, myocardial infarction, and stroke). CORONA, GISSI-HF and AURORA appear to be trying to reproduce the effect observed in JUPITER in conditions where it should be even easier to detect. Why did these replications fail? Perhaps because the decision to interrupt JUPITER for evidence of early benefit was mistaken. (Although it was not exceptional. A systematic review showed that the number of trials that are being stopped early for apparent benefit is gradually increasing (Bassler et al. 2010). It often happens that the decision to stop is not well justified in the ensuing reports (see Nardini 2013): the treatment effects are

often too large to be plausible, given the number of events recorded. Thus, the observed effects are not replicable because researchers ground their conclusions too optimistically on not large enough sample sizes (insufficient power).

Unlike the FDA experts discussed in the previous section, The American College of Cardiology/American Heart Association did not correct for the flaws in JUPITER and we may suspect that they may have been somehow biased by the huge commercial interests at stake. Hence, we need to pay attention not just to the replicability of trials, but also to the way in which experts judge their conclusions.

5.6 Conclusion

We have only covered (partially) the methodological side of the replicability crisis. We have shown how a proper epistemic interpretation of *p*-values requires replicability. This latter depends, on the one hand, on the controls we impose on the experiment to secure that it is not biased by any particular preference or skill of the experimenter (or any other participant in the trial), and, on the other hand, on a proper statistical design for the trial, in which the sample size plays a crucial role. Without a previous agreement on the list of controls and statistical features that characterizes a fair trial, we may be missing replicability due to ambiguity in our experimental plan. And yet, not only statistical replicability matters. As John Norton has recently argued (Norton 2015), the epistemic value of a replication is domain-specific: it depends on what we already knew about a given condition and the goals we seek to reach with a treatment. On the one hand, we need clinically (and not just statistically) significant outcomes. On the other, we need to investigate how experts' judgment can properly assess the statistical evidence provided by trials.

References

Armitage, Peter. 2003. “Fisher, Bradford Hill, and Randomization.” *International Journal of Epidemiology* 32 (6):925–28. https://doi.org/10.1093/ije/dyg286.

Asmar, Roland, and Hassan Hosseini. 2009. “Endpoints in Clinical Trials: Does Evidence Only Originate from ‘Hard’ or Mortality Endpoints?:” *Journal of Hypertension* 27 (Suppl 2):S45–50. https://doi.org/10.1097/01.hjh.0000354521.75074.67.

Bassler, Dirk. 2010. “Stopping Randomized Trials Early for Benefit and Estimation of Treatment EffectsSystematic Review and Meta-Regression Analysis.” *JAMA* 303 (12):1180. https://doi.org/10.1001/jama.2010.310.

Begley, C. Glenn, and Lee M. Ellis. 2012. “Drug Development: Raise Standards for Preclinical Cancer Research.” *Nature* 483 (7391):531–33. https://doi.org/10.1038/483531a.

Bero, Lisa A., and Drummond Rennie. 1996. “Influences on the Quality of Published Drug Studies.” *International Journal of Technology Assessment in Health Care* 12 (2):209–37. https://doi.org/10.1017/S0266462300009582.

Button, Katherine S., John P. A. Ioannidis, Claire Mokrysz, Brian A. Nosek, Jonathan Flint, Emma S. J. Robinson, and Marcus R. Munafò. 2013. “Power Failure: Why Small Sample Size Undermines the Reliability of Neuroscience.” *Nature Reviews Neuroscience* 14 (5):365–76. https://doi.org/10.1038/nrn3475.

Carpenter, Daniel. 2014. *Reputation and Power: Organizational Image and Pharmaceutical Regulation at the FDA*. Princeton University Press.

Carpenter, Daniel, Evan James Zucker, and Jerry Avorn. 2008. “Drug-Review Deadlines and Safety Problems.” *New England Journal of Medicine* 358 (13):1354–61. https://doi.org/10.1056/NEJMsa0706341.

Fellström, Bengt C., Alan G. Jardine, Roland E. Schmieder, Hallvard Holdaas, Kym Bannister, Jaap Beutler, Dong-Wan Chae, et al. 2009. “Rosuvastatin and Cardiovascular Events in Patients Undergoing Hemodialysis.” *New England Journal of Medicine* 360 (14):1395–1407. https://doi.org/10.1056/NEJMoa0810177.

Franklin, Allan. 1990. *Experiment, Right or Wrong*. Cambridge University Press.

GISSI-HF investigators. 2008. “Effect of N-3 Polyunsaturated Fatty Acids in Patients with Chronic Heart Failure (the GISSI-HF Trial): A Randomised, Double-Blind, Placebo-Controlled Trial.” *The Lancet* 372 (9645):1223–30. https://doi.org/10.1016/S0140-6736(08)61239-8.

Goldacre, Ben. 2012. *Bad Pharma: How Drug Companies Mislead Doctors and Harm Patients*. London: Fourth Estate.

González-Moreno, María, Cristian Saborido, and David Teira. 2015. “Disease-Mongering through Clinical Trials.” *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences* 51 (Supplement C):11–18. https://doi.org/10.1016/j.shpsc.2015.02.007.

Gøtzsche, Peter C. 2013. *Deadly Medicines and Organised Crime: How Big Pharma Has Corrupted Healthcare*. Radcliffe Publishing.

Hackshaw, Allan. 2011. *A Concise Guide to Clinical Trials*. John Wiley & Sons.

Helgesson, Claes-Fredrik. 2010. “From Dirty Data to Credible Scientific Evidence : Some Practices Used to Clean Data in Large Randomised Clinical Trials.” In , 49–66. Ashgate. http://www.diva-portal.org/smash/record.jsf?pid=diva2:370359.

Horton, Richard. 2012. “Offline: A Serious Regulatory Failure, with Urgent Implications.” *The Lancet* 379 (9811):106. https://doi.org/10.1016/S0140-6736(12)60032-4.

Hróbjartsson, Asbjørn, Ann Sofia Skou Thomsen, Frida Emanuelsson, Britta Tendal, Jeppe Vejlgaard Rasmussen, Jørgen Hilden, Isabelle Boutron, Philippe Ravaud, and Stig Brorson. 2014. “Observer Bias in Randomized Clinical Trials with Time-to-Event Outcomes: Systematic Review of Trials with Both Blinded and Non-Blinded Outcome Assessors.” *International Journal of Epidemiology* 43 (3):937–48. https://doi.org/10.1093/ije/dyt270.

Ioannidis, J. P. A. 2014. “Clinical Trials: What a Waste.” *BMJ* 349 (dec10 14):g7089–g7089. https://doi.org/10.1136/bmj.g7089.

Ioannidis, John P. A. 2005a. “Contradicted and Initially Stronger Effects in Highly Cited Clinical Research.” *JAMA* 294 (2):218. https://doi.org/10.1001/jama.294.2.218.

———. 2005b. “Why Most Published Research Findings Are False.” *PLOS Medicine* 2 (8):e124. https://doi.org/10.1371/journal.pmed.0020124.

———. 2008. “Why Most Discovered True Associations Are Inflated:” *Epidemiology* 19 (5):640–48. https://doi.org/10.1097/EDE.0b013e31818131e7.

———. 2014. “More Than a Billion People Taking Statins?: Potential Implications of the New Cardiovascular Guidelines.” *JAMA* 311 (5):463–64. https://doi.org/10.1001/jama.2013.284657.

Kahan, Brennan C., Sunita Rehal, and Suzie Cro. 2015. “Risk of Selection Bias in Randomised Trials.” *Trials* 16 (1). https://doi.org/10.1186/s13063-015-0920-x.

Kjekshus, John, Eduard Apetrei, Vivencio Barrios, Michael Böhm, John G.F. Cleland, Jan H. Cornel, Peter Dunselman, et al. 2007. “Rosuvastatin in Older Patients with Systolic Heart Failure.” *New England Journal of Medicine* 357 (22):2248–61. https://doi.org/10.1056/NEJMoa0706201.

La Caze, Adam. 2016. *The Randomized Controlled Trial: Internal and External Validity*. Routledge Handbooks Online. https://doi.org/10.4324/9781315720739.ch18.

Lehrer, Jonah. 2010. “The Truth Wears Off.” *The New Yorker*, December 6, 2010. https://www.newyorker.com/magazine/2010/12/13/the-truth-wears-off.

Lorgeril, Michel de, Patricia Salen, John Abramson, Sylvie Dodin, Tomohito Hamazaki, Willy Kostucki, Harumi Okuyama, Bruno Pavy, and Mikael Rabaeus. 2010. “Cholesterol Lowering, Cardiovascular Diseases, and the Rosuvastatin-JUPITER Controversy: A Critical Reappraisal.” *Archives of Internal Medicine* 170 (12):1032–36. https://doi.org/10.1001/archinternmed.2010.184.

Macedo, Ana Filipa, Fiona Claire Taylor, Juan P. Casas, Alma Adler, David Prieto-Merino, and Shah Ebrahim. 2014. “Unintended Effects of Statins from Observational Studies in the General Population: Systematic Review and Meta-Analysis.” *BMC Medicine* 12 (March):51. https://doi.org/10.1186/1741-7015-12-51.

Marks, Harry M. 1997. *The Progress of Experiment: Science and Therapeutic Reform in the United States, 1900-1990*. Cambridge History of Medicine. Cambridge [England] ; New York: Cambridge University Press.

Matthews, J. Rosser. 1995. *Quantification and the Quest for Medical Certainty*. Princeton University Press.

Mora, Samia, and Paul M. Ridker. 2006. “Justification for the Use of Statins in Primary Prevention: An Intervention Trial Evaluating Rosuvastatin (JUPITER)—Can C-Reactive

Protein Be Used to Target Statin Therapy in Primary Prevention?" *The American Journal of Cardiology*, A Symposium: The Interplay of Dyslipidemia and Inflammation: Reducing Cardiovascular Risk in Diverse Patient Populations, 97 (2, Supplement 1):33–41. https://doi.org/10.1016/j.amjcard.2005.11.014.

Nardini, Cecilia. 2013. "Monitoring in Clinical Trials: Benefit or Bias?" *Theoretical Medicine and Bioethics* 34 (4):259–74. https://doi.org/10.1007/s11017-013-9264-2.

———. 2016. *Bayesian Versus Frequentist Clinical Trials*. Routledge Handbooks Online. https://doi.org/10.4324/9781315720739.ch21.

Norton, John D. 2015. "Replicability of Experiment." *THEORIA. An International Journal for Theory, History and Foundations of Science* 30 (2):229. https://doi.org/10.1387/theoria.12691.

Pereira, Tiago V., Ralph I. Horwitz, and John P. A. Ioannidis. 2012. "Empirical Evaluation of Very Large Treatment Effects of Medical Interventions." *JAMA* 308 (16):1676. https://doi.org/10.1001/jama.2012.13444.

Sorić, Branko. 1989. "Statistical 'Discoveries' and Effect-Size Estimation." *Journal of the American Statistical Association* 84 (406):608–10. https://doi.org/10.1080/01621459.1989.10478811.

Spanos, Aris, and Deborah G. Mayo. 2015. "Error Statistical Modeling and Inference: Where Methodology Meets Ontology." *Synthese* 192 (11):3533–55. https://doi.org/10.1007/s11229-015-0744-y.

Stegenga, Jacob. 2016. *Measuring Harms*. Routledge Handbooks Online. https://doi.org/10.4324/9781315720739.ch31.

Teira, David. 2011. "Bayesian Versus Frequentist Clinical Trials." In *Philosophy of Medicine [Handbook of Philosophy of Science, Vol. 16],* edited by Gifford Fred. Elsevier.

Urfalino, Philippe. 2012. "Reasons and Preferences in Medicine Evaluation Committees." *Collective Wisdom: Principles and Mechanisms. Cambridge, MA*.

CONCLUDING REMARKS

This thesis has been organized around two overlapping questions: (1) whether and how the current approach to the regulation of new medicines should be modified to manage issues generated by new biomedical products, and (2) whether this issue may be reduced to an epistemological and methodological analyses. With regard to this, let us summarize what we have achieved so far.

At first, we have explored the history of the current regulatory system in order to understand why regulators adopted RCTs as the gold standard for testing drugs before granting market approval. When the drug trade emerged at the beginning of the 20th century, quality controls of drugs manufacturing were adopted as the sole standard for drug regulation. In the following years though, a series of *pharmaceutical catastrophes* made clear that neither laboratory test nor experts' judgment provided enough evidence to protect consumers from serious adverse effects. In the same years, innovations in the design of experiments and statistics emerged in agricultural sciences, and gave the opportunity to the medical community to conceive and conduct methodologically sound comparative experiments to test treatments. Despite some early successful applications of the new experimental design (RCTs) in the context of medical testing, FDA had not neither the political nor scientific authority to set new standards until another tragedy occurred in 1962. The notorious thalidomide scandal forced regulators to revise their previous decisions, and to finally implement RCTs as the gold standard for drug testing. This brief historical analysis clearly shows that political considerations, rather than epistemic, were at the basis of the evolution of clinical trials.

Secondly, we have highlighted some practical challenges that nowadays are hindering the conduct of RCTs to test new cancer treatments in the light of the post-genomics revolution. To overcome those difficulties, novel experimental designs are under investigation, adaptive trials being the most promising. However, regulatory agencies have not fully endorsed them yet mainly for two reasons: they are afraid that a lack of standardization could actually transform the decision-making approval into a never-ending process, and there is not a clear scientific consensus on the foundations of novel trial designs. With this regard, we proposed a way to avoid a regulatory impasse while dispensing with a strict regulatory standard, and to facilitate scientific agreement. In a *casuistic* regulatory framework, experts should agree not on *theories* behind the features of every single trial design, but only on single *real cases*, which is far easier. Once regulators have found an agreement on a small set of paradigmatic cases, they can indeed easily assess new ones by analogical reasoning. With regard to this, we suggested as a potential paradigmatic case the clinical development of crizotinib, a molecularly targeted drug for the treatment of lung cancer recently approved on the basis of two small single-arm early phase trials.

In third place, we explored the consequences of a more flexible regulatory system adopting the distinction between rules versus standards as a theoretical framework to capture the costs at stake in the current controversies on regulatory evidence. There are, indeed, good a priori arguments about the superiority of *standards* of evidence (i.e. evidential pluralism) in the detection of drugs safety and efficacy. And there are equally good a priori proposals on how to organize *impartial committees* to carry out such deliberation. However, framing the decision in terms of the costs of implementing these ideals allows us to grasp also the actual trade-off that regulatory authorities should make: there is only a limited amount of resources for them to use. It is possible that the current system based on RCTs have offered so far, a cost-effective

way of protecting patients, if their safety is defined in terms of market withdrawals. In any case, our regulatory experience provides an empirical benchmark for other alternatives. We suggest that, rather than further a priori discussions, we need instead to test these alternatives in a pilot committee.

Fourthly, we have focused on the debate ushered by the recent enactment of the 21st Century Cures Act, which has paved the way for a more flexible drug regulation. Most of its critics are worried by the potential negative impact that it would have on patients. We argued that the multiplicity of testing standards is more defensible than critics think, showing that it is already pervasive in medicine, and that we accept it because we are willing to accept the consequences of potential mistakes. We have argued that the reason for the regulators to consider different testing standards is neither epistemic nor methodological but rather political, and lies in the fact that some medical products are intuitively safer than others and therefore it is less likely that they will produce a medical disaster. In an environment with limited resources, this approach sounds rational. With regard to this, we have suggested to adopt a distinction between *hazards* and *risks* in order to better understand the risk regulators care about. This latter can be defined as the likelihood of a treatment of producing a medical catastrophe. In this context, the size and the definition of target population play a key role. With the progress of biological sciences, nowadays we can circumscribe the target population for certain kind of drugs, such as cancer targeted therapies. In some cases, this population is composed by few hundreds of patients, therefore from a political standpoint targeted therapy do not pose a threat to public health.

In the final chapter, we have covered partially the methodological side of the replicability crisis in the context of clinical trials. We have shown how a proper epistemic interpretation of *p*-values requires replicability. This latter depends, on the one hand, on the controls we impose on the experiment to secure that it is not biased

by any particular preference or skill of the experimenter (or any other participant in the trial), and, on the other hand, on a proper statistical design for the trial, in which the sample size plays a crucial role. Without a previous agreement on the list of controls and statistical features that characterizes a fair trial, we may be missing replicability due to ambiguity in our experimental plan. And yet, not only statistical replicability matters. On the one hand, we need *clinically significant outcomes* (and not just statistically). On the other hand, we need to investigate how experts' judgment can properly assess the statistical evidence provided by trials. Of course, the reliability of evidence is crucial for making sound regulatory decisions, and our analysis shows once again that it does not depend exclusively on the source with originate it, but also on a consensus between experts, which might transcend epistemological considerations.

Coming back to our central questions (1) and (2) we can draw some general conclusions. First of all, there is no way to tackle the issue of changing drugs regulation exclusively from an epistemological perspective. Epistemological and methodological considerations are important, but nonetheless the financial and political issues at stake in drug regulation are enormous. Thus, any epistemological analysis taken out of context cannot make any addition to current debates. Secondly, we must be aware that scientific landscape is rapidly and constantly changing. The molecular revolution in biomedicine makes urgent and crucial to reconsider a regulatory system that has been put in place to regulate products which were essentially different. We should as well take very seriously the fears of the potential negative consequences of adopting looser regulatory standards of evidence, especially because so far RCTs have offered a reasonable level of safety to patients. The good news, in our view, is that there is room for changes without losing much safety for population. The bad news is that changes for the better would require a more systematic and collective effort, which might be beyond any political and academic willingness.

The work collected in this thesis should be understood as aimed at paving the way to the possibility of developing a more rational and coherent *philosophy of drug regulation*. All of the arguments in this thesis therefore add to the growing body of work in the philosophy of medicine, and how it can be applied to the evaluation of medical treatments and technologies

Acknowledgements

Some special thanks go to my supervisors for their patience and support; to all my friends around the campus; and to the former members of FOLSATEC. This program has been for everyone a source of intellectual enlightenment that it could have been hardly found elsewhere. It is regrettable that it has been shut down in such an undignified way.