STATISTICAL EVIDENCE AND THE RELIABILITY OF MEDICAL RESEARCH

Mattia Andreoletti & David Teira

Keywords: Statistical evidence, RCTs, replicability, reliability, false findings.

Statistical evidence is pervasive in medicine. In this chapter we will focus on the reliability of randomized clinical 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*: 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 are big enough.

1. 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. 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 –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 difference, Fisher arranged the experiment as a test of the hypothesis that there is no difference between the two treatments (Teira 2011). This 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 *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 in 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 & 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: see Nardini, this volume. 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 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.

2. 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 2010a). 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, this volume), 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* (see La Caze, this volume): 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 (e.g., Franklin 1990): every experimental setup is different and so are the potential confounding factors and the corresponding controls. Experimenters in all disciplines have their 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 (e.g., 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). 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 (2013) illustrates the risk of unmaskeded 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 (Hrobjartsson *et al*. 2013), 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 (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 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 (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 ("most published research findings are false": see, e.g., Soric 1989, Ioannidis 2005a, 2005b, 2014a), 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 et al. 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 (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 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.

3. 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 et al. (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 instead of *hard* trial outcomes (death); sometimes not. 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-grounded 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 (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 2010a, p. 623). 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 et al. 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, this volume. 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 et al. 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.

4. 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, e.g., Goldacre

2012, González-Moreno et al. 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 (2014b) words, this would amount to a "statinization" of the planet.

Taking statins is not completely harmless: there are side effects (Macedo et al. 2014). So what were the grounds for such a massive public health intervention? According to Ioannidis (2014b), 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) (Ridker 2009). 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 were someone 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 unsuccessfully. For instance, CORONA (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 outcome of CORONA was also reached by two more trials: GISSI-HF (2008) and AURORA (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: 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.

Concluding remarks

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 (2015) has recently argued, 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 expert judgment can properly assess the statistical evidence provided by trials.

Word count: 6007

Related topics

Outcome measurements in medicine (Leah McClimans) - The randomized controlled trial: internal and external validity (Adam LaCaze) - Systematic review, meta-analysis and the evidence hierarchy (Robyn Bluhm) - Bayesian versus frequentist interpretations of clinical trials (Cecilia Nardini) - Subjective and objective probabilities in causal models in medicine. (Donald Gillies) - Translational Medicine (Jason Robert)

References

AURORA STUDY GROUP. 2009. Rosuvastatin and Cardiovascular Events in Patients Undergoing Hemodialysis. *New England Journal of Medicine*. 360:1395-1407.

BASSLER, D., BRIEL, M., MONTORI, V. M., LANE, M., GLASZIOU, P., ZHOU, Q., HEELS-ANSDELL, D., WALTER, S. D., GUYATT, G. H.; STOPIT-2 STUDY GROUP, FLYNN, D. N., ELAMIN, M. B., MURAD, M. H., ABU ELNOUR, N. O., LAMPROPULOS, J. F., SOOD, A., MULLAN, R. J., ERWIN, P. J., BANKHEAD, C. R., PERERA, R., RUIZ CULEBRO, C., YOU, J. J., MULLA, S. M., KAUR, J., NERENBERG, K. A., SCHÜNEMANN, H., COOK, D. J., LUTZ, K., RIBIC, C. M., VALE, N., MALAGA, G., AKL, E. A., FERREIRA-GONZALEZ, I., ALONSO-COELLO, P., URRUTIA, G., KUNZ, R., BUCHER, H. C., NORDMANN, A. J., RAATZ, H., DA SILVA, S. A., TUCHE, F., STRAHM, B., DJULBEGOVIC, B., ADHIKARI, N. K., MILLS, E. J., GWADRY-SRIDHAR, F., KIRPALANI, H., SOARES, H. P., KARANICOLAS, P. J., BURNS, K. E., VANDVIK, P. O., COTO-YGLESIAS, F., CHRISPIM, P. P,, RAMSAY, T. 2010. Stopping randomized trials early for benefit and estimation of treatment effects: systematic review and meta-regression analysis. *JAMA*, 303(12):1180-7.

BEGLEY, C. G., ELLIS, L. M. 2012. Drug development: Raise standards for preclinical cancer research. *Nature*, 483, 531–533.

BERO, L., RENNIE, D. 1996. Influences on the Quality of Published Drug Studies. *International Journal of Technology Assessment in Health Care*, 12.2, 209-237.

BUTTON, K. S., IOANNIDIS, J. P., MOKRYSZ, C., NOSEK, B. A., FLINT, J., ROBINSON, E. S., & MUNAFÒ, M. R. 2013. Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience,14*(5), 365-376.

CARPENTER, D. P. 2010. *Reputation and power : organizational image and pharmaceutical regulation at the FDA*, Princeton, Princeton University Press.

CARPENTER, D., ZUCKER, E. J. & AVORN, J. 2008. Drug-review deadlines and safety problems. *N Engl J Med,* 358**,** 1354-61.

CORONA GROUP. 2007. Rosuvastatin in older patients with systolic heart failure. *New England Journal of Medicine,* 357(22):2248-61.

DE LORGERIL M., SALEN M. P., ABRAMSON J. , DODIN S., HAMAZAKI T., KOSTUCKI W., OKUYAMA, H. PAVY, B., RABAEUS, M. 2010. Cholesterol Lowering, Cardiovascular Diseases, and the Rosuvastatin-JUPITER Controversy. A critical reappraisal. *Arch Intern Med,* 170(12):1032-1036.

FRANKLIN, A. 1990. *Experiment, Right or Wrong*, Cambridge, Cambridge University Press.

GILLIES, D. 2000. *Philosophical Theories of Probability*, London, Routledge.

GISSI-HF INVESTIGATORS. 2008. Effect of rosuvastatin in patients with chronic heart failure (the GISSI-HF trial): a randomised, double-blind, placebo-controlled trial. *The Lancet*, 372(9645):1231-1239.

GOLDACRE, B. 2012. *Bad Pharma*, London, Fourth State.

GONZÁLEZ-MORENO, M., SABORIDO, C., TEIRA, D. 2015. "Disease-mongering through clinical trials," *Studies in History and Philosophy of Biological and Biomedical Sciences*, 51, 11-18.

GREENE, J. A. 2007. *Prescribing by numbers : drugs and the definition of disease,* Baltimore, Johns Hopkins University Press.

GØTZSCHE, P. 2013. *Deadly Medicines and Organised Crime: How big pharma has corrupted healthcare*, London: Radcliffe Health.

HACKSHAW, A. K. 2009. *A concise guide to clinical trials,* Chichester, UK ; Hoboken, NJ, Wiley-Blackwell/BMJ Books.

HELGESSON, C-F. 2013. From dirty data to credible scientific evidence: Some practices used to clean data in large randomised clinical trials. *In*: WILL, C., MOREIRA, T. (eds.) *Medical Proofs, Social Experiments*, Farnham, Ashgate, 2010, pp. 49-64.

HORTON, R., 2015. Offline: What is medicine's 5-sigma?. *The Lance*t, 385(9976):1380.

HROBJARTSSON, A., THOMSEN, A. S., EMANUELSSON, F., TENDAL, B., HILDEN, J., BOUTRON, I., RAVAUD, P. & BRORSON, S. 2013. Observer bias in randomized clinical trials with measurement scale outcomes: a systematic review of trials with both blinded and nonblinded assessors. *Cmaj*, 185**,** E201-11.

LEHRER, J. 2010. The Truth Wears Off. *The New Yorker* [Online]. Available: http://www.newyorker.com/magazine/2010/12/13/the-truth-wears-off [Accessed June 15 2015]

IOANNIDIS, J. P 2005a. Why Most Published Research Findings Are False. *PLoS Med*, 2(8): e124. doi:10.1371/journal.pmed.0020124.

IOANNIDIS, J. P 2005b. Contradicted and Initially Stronger Effects in Highly Cited Clinical Research. *JAMA*, 294(2):218-228.

IOANNIDIS, J. P. 2008. Why most discovered true associations are inflated. *Epidemiology*, 19**,** 640-8.

IOANNIDIS, J. P 2014a. Clinical trials: what a waste. *BMJ*, 349:g7089.

IOANNIDIS, J. P 2014b. More than a billion people taking statins?: potential implications of the new cardiovascular guidelines. JAMA, 311: 463–464.

MACEDO, A. F., TAYLOR, F. C., CASAS, J. P., ADLER, A., PRIETO-MERINO, D., EBRAHIM, S. 2014, Unintended effects of statins from observational studies in the general population: systematic review and meta-analysis, *BMC Medicine*, 12:51.

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

MATTHEWS, J. R., 1995. *Quantification and the Quest for Medical Certainty*, Princeton University Press, Princeton, New Jersey.

NORTON, J. 2015. Replicability of experiment. *Theoria*, 30/2, 229-48.

PEREIRA, T. V., HORWITZ, R. I. & IOANNIDIS, J. P. 2012. Empirical evaluation of very large treatment effects of medical interventions. *Jama*, 308**,** 1676-84.

RIDKER, P. M., COOK, N. R. Statins: new American guidelines for prevention of cardiovascular disease. *The Lancet*, 382(9907):1762-5.

SORIC, B. 1989. Statistical "Discoveries" and Effect-Size Estimation. *Journal of the American Statistical Association*, 86(406):608-610.

SPANOS, A. & MAYO, D. 2015. Error statistical modeling and inference: Where methodology meets ontology, *Synthese* (Online first), 10.1007/s11229-015-0744-y

TEIRA, D. 2011. Frequentist versus Bayesian Clinical Trials. *In:* GIFFORD, F. (ed.) *Philosophy of Medicine.* Amsterdam: Elsevier, pp. 255-297.

URFALINO, P. 2012. Reasons and Preferences in Medicine Evaluation Committees. *In*: ELSTER, J. LANDEMORE, H. (eds.) *Collective Wisdom. Principles and Mechanisms*, Cambridge, Cambridge University Press, pp. 173-203.

Further readings

Since the replicability crisis is still unfolding it is probably better to use the Internet for updated references.

For a general overview, in open access, you will find *Nature*'s special issue on reproducibility: http://www.nature.com/nature/focus/reproducibility/index.html. For updates on withdrawn papers from scientific journals, often (but not only) for replicability issues see http://retractionwatch.com/ Deborah Mayo's blog is a rich source of (statistically informed) philosophical discussion on the replicability crisis: http://errorstatistics.com/category/reproducibility/ An extensive historical source about

controls implemented medical experiments for grounding like with like comparisons is the James Lind library: http://www.jameslindlibrary.org/