

ARTICLE

The Elastic Ruler: Placebo Responses, Clinical Trials, and Medical Knowledge

Shane N. Glackin 

EGENIS/Department of Social and Political Sciences, Philosophy and Anthropology, University of Exeter, Exeter, UK

Email: s.n.glackin@exeter.ac.uk

(Received 10 May 2024; revised 06 September 2024; accepted 09 January 2025; first published online 16 January 2025)

Abstract

This article first poses a skeptical challenge to clinical trials in medicine. The efficacy of treatments is measured against placebo, but placebo responses are not constant. They fluctuate with demographic variables, and they seem to be increasing over time. We therefore find ourselves measuring with the equivalent of what Wittgenstein termed an “elastic ruler.” I then propose a “skeptical solution” to the problem. Elastic rulers are suitable tools for measuring dynamic, floating networks of values, like foreign currency exchanges. We can assuage the skeptical concerns by understanding clinical trials in this way; I suggest several practical guidelines for doing so.

When you can measure what you are speaking about, and express it in numbers, you know something about it; but when you cannot express it in numbers, your knowledge is of a meagre and unsatisfactory kind; it may be the beginning of knowledge, but you have scarcely, in your thoughts, advanced to the stage of *science*, whatever the matter may be.

William Thomson (1889, 73–74, quoted in Boumans 2015, 26)

*

Measurement is a basic instrument of modern thought. We use it to describe the world, to reason about it, and to manage it. We seek strength in numbers. Our culture has, to a large extent adopted the programme, which Galileo laid down for the physical sciences: “Count what is countable, measure what is measurable and make measurable that which is not.”

Ludwik Finkelstein (2008, 1)

*

It might just be condensation. If the dripping was regular then it was just condensation. He would stop and listen, measuring the sound against his heartbeat. It sounded regular. He breathed out. But what if his heart wasn't beating regularly? He would stop breathing and listen, and his wretched heart would begin an irregular beat.

Ben Smith (2019, 37)

1. Introduction

Scientific knowledge requires measurement; measurement requires appropriate instruments. But are our instruments appropriate? An elastic ruler, Wittgenstein tells us (1978, 38–39; see also Reichenbach 1958, 16), would be singularly useless for the purposes of measurement, at least for creatures like us; we would find any purported practice of measurement employing such a tool completely unintelligible. But this, as I shall argue, seems to be the situation of virtually all modern medical knowledge.

I begin by considering the importance of Randomized Controlled Trials (RCTs), commonly referred to as the “gold standard” for measuring the efficacy of medical therapies. These trials compare the apparent therapeutic effect of a candidate treatment against that observed when only a placebo is given; therapeutic response to placebo is therefore the “zero point” against which treatment efficacy is measured. But as I will outline, this zero is not constant; it varies with a host of demographic factors—such as culture, age, sex, nationality, education-level—as well as genetics, and it appears to be strengthening across all such groups, albeit in some countries only. Medicine, in short, is measured with an elastic ruler.

Following Stevens's classic schema (1946), Measurement Theory distinguishes between four kinds of scale: nominal, ordinal, interval and ratio. I show that, while medical intervention is usually treated as though its empirical structure corresponds (“maps”) to a ratio or an interval scale, in fact it maps at best to an ordinal scale—which has been argued not to properly qualify as “measurement” at all (Campbell 1928)—and perhaps even to something weaker again. Put another way, if placebo-controlled trials are an *instrument* for measuring medical efficacy, I argue, we have no way of calibrating it.

Fortunately, all is not lost. Taking the ubiquitous analogy with the “gold standard” seriously, and drawing on Peter Hacker's (2015, 17ff.) discussion of currency exchanges as an example of a functioning system of elastic rulers, I propose a “skeptical solution” (following Kripke 1982) whereby we accept the lack of an objective measure for medical efficacy, but show that we can nevertheless do without one. I close by examining how exactly to “do without one,” suggesting some necessary reforms to the clinical trials process and scholarly and popular reporting protocols that follow from reconceiving it in this light.

2. RCTs: The “gold standard”

There are countless ways in which we typically wish to improve the world: raise per-capita GDP, reduce the prison population, cure somebody's leukemia. And there are probably even more ways that we might propose to intervene in the world to bring about those improvements: tax-cuts, investment in infrastructure, education programs, harsher deterrentist sentencing, chemotherapy, prayer . . . The problem is, we can't do everything. Some interventions are useless or actively exacerbate the

problem, like cutting government spending in a recession. Some are mutually exclusive; we cannot treat the same ailment simultaneously with bed-rest and a course of vigorous exercise. And some are simply costly in terms of finite resources; we can't afford to try everything proposed. So to justify interventions, we need some way of knowing how effective they actually are.

The theoretically ideal way to do this would be to make a counterfactual comparison between the world as it would be both with and without the intervention in question; how does this patient's leukemia progress in the counterfactual world where we pray for remission, compared to the one in which we don't? But this raises what is known as the "fundamental identification problem" (Holland 1986; Rubin 1974, 2005); counterfactual worlds are epistemically inaccessible, so we cannot simultaneously observe the same subject both with and without the intervention.

Various ways of surmounting the fundamental identification problem are possible, but across the social sciences—in which we may for current purposes include medicine—the highest evidentiary value for inferring the causal influence of interventions is accorded to the Randomized Controlled Trial. This is referred to, with striking ubiquity, as the "gold standard" for such evidence; at the time of writing, a Google Scholar search for "RCT gold standard" yields about 214,000 results. And the image is sometimes extended; one writer, for instance, describes "the very highest carat evidential gold" (Howick 2017, 1364). I will return to this—unintentionally revealing—metaphor later in the discussion.

Because we cannot observe the same subject, at the same time, both with and without intervention, RCTs aim to "replace the estimation of individual treatment effects with average treatment effects" (Paul and Healy 2018, 324). That is, such trials randomly and blindly assign a study population into treatment (intervention) and control (no intervention) groups. This does not represent a "true" experimental control, where all potential variables can be matched and adjusted; instead, blind randomization functions as a proxy for such adjustment by "evening out" individuals' differences between the groups. This means that the heterogeneity of individual study participants can for practical purposes be disregarded, since the study now estimates the *average* effects of intervention and nonintervention across groups that are as close as possible to equivalent. And while any trial result may still be confounded by subjects' idiosyncrasies, the RCT methodology allows us to statistically estimate the probability that this has taken place, and therefore announce with precision our degree of warranted confidence in the finding (though see Deaton and Cartwright 2018 for a skeptical view of several parts of this summary).

Where trials of *medical* interventions—i.e. Phase II/III clinical trials—are concerned, the control group is usually¹ given a placebo treatment, rather than no treatment at all. The efficacy of any treatment is assessed, straightforwardly enough, by comparing outcomes between treated and untreated groups. The particular issue in medical trials is that the act of treating, by itself, seems to involve, or induce, any number of physiologically irrelevant variables that can nevertheless affect the outcome. So to assess the efficacy of a medical intervention we need to know whether,

¹ This claim is, strictly speaking, inaccurate; often they are given the standard existing treatment instead. This muddies the waters of my argument a little, but only a little, so I will proceed as though placebo controls are ubiquitous, and come back to the point in section 5.

and to what extent, it outperforms *the generic effects of treatment*, rather than the effect of leaving the condition untreated. The control group is therefore provided with a treatment which is as far as possible indistinguishable—including, where feasible, to those conducting the trials—from that given to the treatment group, except with regard to the specific variable the trial assesses: an identically administered but pharmacologically inert substitute for a new drug, for instance.

The key point to take from this is: *medical efficacy is therefore defined relative to placebo, not in absolute terms*. When we ask if a treatment “works,” we are not asking if you will probably get better; you probably will, whether or not you take the treatment, since the body has evolved to be able to fight most ailments it encounters.² We are asking if it works any better than not taking the treatment. And the proxy we use for “not taking the treatment”, since *the mere act of treating* can affect the outcome, is double-blind treatment with placebo.

3. Instability of the placebo response

There is some controversy about what, if anything, “the placebo effect” is (e.g., Shapiro 1968; Grünbaum 1986; Gotzsche 1995; Louhiala et al. 2013; Howick 2017). I will sidestep those issues here; rigorous conceptual clarity on this point is not really relevant to the argument I want to make, while the existence of the effect is at least very widely attested, and unless it can be categorically disproved must anyway be controlled for in clinical trials. It can be helpfully, if whimsically, summarized as “the healing power of a white coat and stethoscope”; under appropriate conditions, patients’ conditions are reported to improve just as a result of exposure to the trappings of medical institutions and practices. And this is not, or not straightforwardly, a matter of “trickery”; placebo responses are observed even in cases where the patient knows that they are receiving a placebo, and what that means (e.g., Park and Covi 1965; Aulas and Rosner 2003).

Among many other colorful facts in the placebo literature, one of the most striking is that the strength of the placebo response is not constant, but rather is known to be strongly influenced by a whole host of apparently irrelevant demographic and contextual factors. A classic paper by Daniel Moerman (2000) observed that placebo responses vary according to factors like physician enthusiasm, tablet color, presence of a brand name, whether a patient (or in the case of children, their mother) had spoken to doctors or nurses, patients’ race and commitment to the tenets of their culture, and how strictly the patients comply with their placebo regime. A more recent literature survey finds evidence that response strength is influenced by “the perceived complexity of the intervention . . . the perceived cost of the procedure, colour of the pill, presence of other patients, competence and warmth of the healthcare provider, and expectations of the patient” (J. A. Olson et al. 2021, 1–2). Results keep coming nearly as quickly as new medical advances; a 2023 study suggests that responses are stronger when the placebo is presented as an instance of “personalised” or “precision” medicine (Sandra et al. 2023). Placebo responses are stronger among women (E. M. Olson et al. 2021, though see Weimer et al. 2015), the

² Or, if the ailment is one that your body cannot fight, you probably won’t, whether or not you take the treatment.

young (Weimer et al. 2015), and white people (Okusogu et al. 2020). Those with higher levels of education are more likely to report negative side-effects from placebo treatments (Bavbek et al. 2015). Moreover, the effect appears to be markedly increasing over time, at least in the United States, but not (or not at the same rate) in other countries (Tuttle et al. 2015).

This is, as I say, a striking series of findings. And it is one that poses an unexpected problem for RCTs. In short, if medical efficacy is relative to placebo, and placebo responses are in all sorts of respects variable, then there is no stable way to establish or assess the efficacy of a given medical treatment. Placebos were supposed to mark the zero-point from which measurement of treatment efficacy began, but if “zero” is no longer a constant, then efficacy cannot be meaningfully measured.

The general form of this problem (and perhaps its solution) was noted by Ludwig Wittgenstein, supposedly while watching the 1936 Eddie Cantor film *Strike Me Pink* in which a dishonest shopkeeper uses an elastic ruler to cheat his customers, stretching it while measuring out the cloth in front of them, then relaxing it when cutting so as to give them less than they were charged for (Rhees 1970, 121–22; cited in Schroeder 2015, 115).³ He developed the thought further at the beginning of his posthumously published *Remarks on the Foundation of Mathematics* (1978):

How should we get into conflict with truth, if our foot rules were made of very soft rubber instead of wood and steel? . . . [W]e should not get, or could not be sure of getting, that measurement which we get with our rigid rulers. So if you had measured the table with the elastic rulers and said it measured five feet by our usual way of measuring, you would be wrong . . . If a ruler expanded to an extraordinary extent when slightly heated, we should say—in normal circumstances—that that made it unusable. (1953, I §5)

“But surely,” Wittgenstein’s interlocutor objects, “that isn’t measuring at all!” (ibid.) Crispin Wright elaborates on this reaction:

It is a feature of the concept of measuring that an accurately measured object will yield distinct readings at distinct times only if it changes; so much is implicit in the notion that measuring is to ascertain a property of the object measured. (1980, 58)

³ Unfortunately for the historical record, and the accuracy of either Rhees’s or Wittgenstein’s recollection, no such episode occurs in the version of *Strike Me Pink* that I have seen (which is of admittedly dubious provenance). Cantor does play a tailor/shopkeeper, but his character is a decided nebbish who gains confidence with the aid of a self-help book, and in no way dishonest. I believe a more likely inspiration is Cantor’s earlier “Moe the Tailor” vaudeville sketch with Lou Hearn and Louis Sorin in *Glorifying the American Girl* (1929); although no elastic ruler features, a variety of similar measurement-related swindles—such as standing the customer on a box and measuring leg-length to the floor—are attempted by the unscrupulous Moe. Cantor also plays a tailor in the silent *Kid Boots* (1926) but the character is, again, hapless rather than crooked, and there is no elastic ruler. Ten solid hours of research, this footnote took.

Here is the worry about placebo response instability in essence, then; when we purport to measure the effect of medical treatments by reference to placebo, then if the placebo response is not constant we effectively do so with an elastic ruler. Our scale of measurement stretches and shrinks, for reasons unconnected to any change in what we are trying to measure. “And surely,” to quote a phrase, “that isn’t measuring at all.”

4. The view from measurement theory

Almost uniquely among major philosophers, Wittgenstein’s primary training was as an engineer, and we can make the worry here more precise using resources from the somewhat neglected subfield of philosophy of engineering. According to a foundational text in measurement theory (Stevens 1946), there are broadly four kinds of measurement scale (see Tal 2013, 1164):

Nominal: A nominal scale is one containing distinct categories that are not ordered; a typical example is the division of rocks into igneous, sedimentary, and metamorphic types.

Ordinal: An ordinal scale is one that is sufficiently ordered to allow ranking and comparison, but not meaningful mathematical operations. The Beaufort Wind Scale, for instance, permits ranking—any Force 5 wind is stronger than any Force 4—but we do not get a Force 4 wind by doubling the speed of a Force 2, or adding the speeds of a Force 1 and Force 3. A nonscientific example would be the scoring system in tennis; a score of 40 is greater than a score of 15, but it makes no sense to say that it is 2.67 times greater.

Interval: An interval scale has an arbitrary zero-point, meaning that while arithmetical operations can be performed on the intervals between values, such operations are not possible on the values themselves. The Celsius and Fahrenheit temperature scales exemplify this; the zero-point of each is defined by an objective physical measure (the freezing-point of water at 1013 hPa, and that of a salt/water brine concocted by Daniel Gabriel Fahrenheit himself,⁴ respectively), but these are set by human convenience and interests rather than any intrinsic feature of the physical property being measured. Accordingly, while 10°C is not “twice as hot” as 5°C, an increase of 10°C is twice as much of an increase as an increase of 5°C.

Ratio: A ratio scale is one in which zero is non-arbitrary. By contrast to Celsius and Fahrenheit, the Kelvin temperature scale has as its zero-point the coldest temperature permitted by the laws of thermodynamics, at which molecular activity in principle ceases. Most physical quantities—of time, distance, mass, volume, etc.—possess a genuine non-arbitrary zero-point in this way, and so display the same algebraic structure as the real numbers; 40km/h is twice as fast as 20km/h, while a 56Ω and a 220Ω resistor placed in series will produce a total resistance of 276Ω.

What can we say about placebo-relative measurements in this light? We often talk as through trial results map to a ratio scale: “such-and-such a treatment is twice as effective as its competitors” and so forth. And this seems natural, because very often we appear to be dealing with straightforward physical quantities, e.g., 5μg of toxin per ml of blood. But measurement in medicine, as in the social and

⁴ I am indebted to an anonymous reviewer for this excellent piece of trivia.

psychological sciences, is much more theoretically (if not always practically) complicated than in the physical sciences. What we are truly interested in when we measure medical efficacy is not micrograms per liter over zero, but *over placebo*. The quantity $5\mu\text{g/ml}$ may be fixed and objective, but $5\mu\text{g/ml}$ minus a placebo value whose magnitude varies between 3 and $4\mu\text{g/ml}$ depending on conditions of measurement is not!⁵

In theoretical terms we might defend the placebo as a non-arbitrary zero-point; in practice, it is worse than a merely arbitrary one because it is inconstant as well.⁶ This leaves us with something rather weaker than an interval scale; operations on intervals seem to be possible, at least intra-trial (e.g., if one treatment reduces mortality by 4 percent relative to the placebo wing, and another reduces it by 6 percent relative to *the same placebo wing*, we might fairly say that the second one has done 50 percent better), but between different trials it may not even allow us to make the consistent transitive orderings typical of ordinal scales. That is, if Treatment A outperforms placebo by 10 percent in Trial A, and Treatment B outperforms placebo by only 5 percent in Trial B, we seemingly have no basis to say that Treatment A is better than Treatment B. It is controversial that even an ordinal scale counts as measurement: “(f)or Campbell (1938), a necessary condition for measurement was that the manifestation of properties or attributes should be additive” (Finkelstein and Leaning 1984, 25). But if not even transitive orderings are possible, “then it can be argued that the attribute concept has no empirical significance (Campbell 1928) and is redundant” (Finkelstein and Leaning 1984, 27). It becomes hard to see how we can draw any reliable conclusions at all about the efficacy of medical treatments; if placebo-controlled trials are the instrument for measuring clinical efficacy, then the instrument looks impossible to calibrate.

5. Objections and responses

Before I try to rehabilitate placebo-controlled efficacy trials, it is worth considering some alternative objections to the skeptical analysis I have sketched. The first of these was mentioned earlier; many RCTs are conducted not with a placebo-controlled arm, but using the existing standard treatment as a control. It is difficult to estimate overall proportions, but of the top ten most widely-cited RCTs considered in Krauss (2018, 318–19), four were placebo-only, five were conventional-treatment-only, and one used both.

There are various reasons for using active-control rather than placebo-control trials. The chief one is ethical; giving trial subjects a placebo when an effective treatment is known to be available arguably violates the principle of clinical equipoise, meaning that placebo controls are ethical only in first-generation trials

⁵ But given that on any particular trial occasion the average placebo effect will be a specific value—say $3.754\mu\text{g/ml}$ —and that a well-designed trial will elicit the same placebo effects in both the active and control arms, can't we simply deduce the specific effect by subtracting the control result from the active one? Unfortunately, this will only allow us to state the effect observed *on that particular occasion*; to generalize further would require an assumption of what is called “additivity” of placebo and specific effects which, as we shall see, is empirically doubtful.

⁶ It is, we might say, arbitrary both *de dicto* and *de re*. For more on the importance of constants in establishing scientific measures, and the question of arbitrariness, see Riordan (2015).

(e.g., Hill 1963; Freedman 1987, 1–6), though others (e.g., Miller and Brody 2003) disagree (see Anderson 2006, 66–68). The World Medical Association's *Declaration of Helsinki* (World Medical Association 2013) seem to give a general, though vague and not exceptionless, injunction against such trials, while others have argued that the results of such trials should not even be considered for publication (Rothman and Michels 1994, as cited in Quitkin 1999).

For present purposes, the ethical issues need not concern us. The question here is whether the existence, and perhaps prevalence, of active-control trials allows us to assess medical efficacy without running into the “elastic ruler” problem that, I have been arguing, casts a pall of doubt on the results of placebo-control ones. I think that it does not.

The reason can be briefly stated; unless we have a measure of the standard existing treatment's efficacy that we know to be stable, the standard existing treatment does not provide a good benchmark either. But of course, if first-generation trials must be performed against placebo given the lack of alternative treatments, the whole problem is that we cannot have a stable measure of those treatments' efficacy. And it follows that we therefore cannot have a stable measure of efficacy for subsequent generations of treatments that are benchmarked against them or their successors; any network of active-control trials will ultimately have its foundations in the shifting sands of placebo control.⁷

A second and potentially more troublesome objection will occur at this point, if it hasn't already, to keen readers of the history of science. In Hasok Chang's seminal *Inventing Temperature* (2004), he tells the story of the development of reliable thermometers, which at first faced a familiar-sounding problem: since even water did not boil and freeze at the same temperatures in all conditions,

there were no standard “fixed points,” namely phenomena that could be used as thermometric benchmarks because they were known to take place always at the same temperature. Without credible fixed points it was impossible to create any meaningful temperature scale, and without shared fixed points used by all makers of thermometers there was little hope of making a standardized scale. (ibid., 9)

The solution, to cut a long and fascinating story short, was a process of what Chang describes as “iterative self-correction”; even though thermometer-makers started out with no reliable, non-ordinal measurements (one could tell by using the same thermometer whether one thing was hotter than another, or had heated or cooled over time, but little else), it was possible to gradually improve both measuring apparatuses and benchmarks by continually recalibrating each in turn against the better and better measurements which resulted. Over time, therefore, temperature measurement could pull itself up by the proverbial bootstraps to the astonishing standards of precision and accuracy that we have today.

⁷ This is closely related to what is sometimes known as the “assay sensitivity” argument, as articulated in the widely adopted ICH E10 Guidelines (2000), though that argument is concerned with the external validation of controls, rather than with their variability. For a positive assessment of the argument see Temple and Ellenberg (2000), and for a negative one see Anderson (2006).

So this raises the question; why can't the myriad of *n*th-generation active-control trials be understood as part of just this kind of process of iterative self-correction? On this view, the fact that the foundational measurements of treatment efficacy are based on unstable placebos rather than fixed points won't matter because each subsequent trial refines and improves on our previous estimations.

There is something right about this view of the collective mass of trials as forming a self-supporting network, and the solution to the problem that I pose in the following section will work in at least a superficially similar way. But a straightforward defence of efficacy measures, based on an analogy between iterative self-correction in the development of other measuring apparatuses, and a combination of placebo-control and 2nd-and-subsequent-generation active-control trials, will not work. The problem is that the iterative self-correction described by Chang drives convergence between techniques and devices of measurement and the object of measurement *because* that object is a property of independent physical reality; we can iterate our way to accurate and reliable measurements of temperature because temperature—the average kinetic energy of the atoms in a substance—is an objective property on which our practices of measurement can converge. It would seem at best question-begging to suppose that medical efficacy is like this, when the same RCT methodology we rely on to measure it is typically used in social sciences to assess properties that are assuredly not.

Indeed, we have positive reason to think that medical efficacy is not like this, and that placebo-control or active-control trials cannot be iteratively calibrated in this way. Throughout the article I have made the simplifying assumption that placebo effects and treatment effects are *additive*; that is, that the total curative effect of any treatment can be partitioned into effects resulting just from the act of treating the patient (placebo effects) and effects arising from the particular treatment administered (specific effects), and that the efficacy of the treatment can therefore be calculated by taking the total curative effect as averaged over the active arm of the trial, and subtracting the average result in the placebo-only wing (see e.g., Meissner et al. 2011; Ho 2023, 7).

This has been a central assumption of the theoretical literature since the modern study of placebo responses began (Beecher 1955). As early as 1960, though, it was observed that “interaction effects” between placebo and treatment can result in total effects that are either more or less than the sum of the two effects (Modell and Garrett 1960); a growing number of studies and surveys of interaction effects in recent years has cast considerable doubt on whether the additive model is tenable (see e.g., Boussageon et al. 2008; Boussageon et. al 2022; Hall and Loscalzo 2019; Kube and Rief 2017). While I have assumed additivity for ease of exposition, the “elastic ruler” criticism of RCT methodology in the biomedical sciences that I have been developing here does not depend on that assumption, because the epistemic problem posed by the variable contributions of placebos to a known total effect is if anything exacerbated by the existence of interaction effects. Nor does the solution I develop in the next section require the assumption of additivity, though I will continue to talk in additive terms for simplicity's sake.

However, the rejection of additivity looks potentially fatal to any defence of RCT methodology based on iterative self-correction. What that putative defense claimed, recall, was that an objective, independent feature of the universe—the efficacy of

some given treatment—could be isolated and quantified despite the variability of the placebo effects that we inevitably measured along with it and the consequent lack of a “fixed point” to measure from. But if additivity does not hold, then there is no objective, independent feature of the universe to be measured; the efficacy of treatment A over and above treatment B will vary according to how each interacts with whatever placebo effects accompany them. And even in an active-control trial, placebo effects will always accompany them; placebo effects are after all those healing effects that result from giving just any treatment at all, regardless of the treatment’s specific effects. In short, the failure of the additive assumption means that iterative self-correction in the absence of fixed points is not possible for measures of medical efficacy; it would not be possible even if placebo effects were constant.

6. The “gold standard” again: A skeptical solution

We have already phrased our skepticism about measurement of medical treatment’s efficacy in Wittgensteinian terms; our only means for achieving it in effect “measures with an elastic ruler,” which surely, we are inclined to say along with Wittgenstein’s interlocutor, “isn’t measuring at all.” To resolve the problem, we will return to Wittgenstein.

In an influential reading of Wittgenstein’s *Philosophical Investigations* (1953), Saul Kripke (1982) attributes to him a strategy he says originates with, and is first named by, David Hume;

In his *Enquiry* (1748/1993), after he has developed his “Sceptical Doubts Concerning the Operations of the Understanding,” Hume gives his “Sceptical Solution of These Doubts.” What is a “sceptical solution”? Call a proposed solution to a sceptical problem a *straight* solution if it shows that on further examination the scepticism proves to be unwarranted . . . A *sceptical* solution of a sceptical philosophical problem begins on the contrary by conceding that the sceptic’s negative assertions are unanswerable. Nevertheless our ordinary practice or belief is justified—contrary appearances notwithstanding—it need not require the justification the sceptic has shown to be untenable.

A skeptical solution to our current skeptical problem, then, will be one that accepts that what we thought we needed—a fixed zero-point from which to derive stable measures of treatment efficacy—is something that we cannot have, but argues that we can nevertheless do without it.

What is meant by describing RCTs—as so many do—as the “gold standard” of medical evidence? Usually that they represent the best possible evidence from which to infer a causal relationship between treatment and outcome, and that they are therefore always preferable to other methods of study except where practical or ethical considerations preclude them (we would not countenance, for instance, studying the developmental effects of childhood malnutrition by deliberately feeding a control group of infants an inadequate diet, though we can of course study by other means the effects of such a diet on those children our society assigns it to for all sorts of other reasons). But even those deeply critical of attributing RCTs that status (e.g., Cartwright 2007) rarely examine the “gold standard” metaphor.

One notable exception to this is Jones and Podolsky (2015), who trace the first use of the phrase in this connection to Feinstein and Horwitz (1982). As they note, the term originates in monetary policy, beginning with Isaac Newton's tenure as Master of the Royal Mint in 1717 and spreading around the world under British trading influence until it began to collapse at the time of the Great Depression; in 1971, the United States moved the dollar—to which most other international currencies were then pegged under the Bretton Woods Agreement—off the gold standard, marking its general demise.

Under the gold standard, each unit of national currency was in principle exchangeable for a set quantity of gold from the nation's reserves (Hobsbawm 1995, 95). According to modern-day supporters of a return to the gold standard, requiring currency to be "backed by" or redeemable for a fixed amount of gold makes it uniquely resistant to inflation and to the effects of government economic policy, compared to so-called fiat money; when money supply is limited by gold reserves it can be increased only at significant cost, whereas "[t]he printing press, on the other hand, is inexhaustible and works like a stroke of magic" (Marx 1993, 121).

To call randomized control tests a "gold standard," then, is to imply that they are particularly trustworthy compared to any alternative; and whether consciously or not, it is to do so on the specific grounds that they are grounded in something fixed and objective, not susceptible to manipulation. But of course, in economic terms this is the purest fantasy; during the quixotic presidential campaign of gold enthusiast Ron Paul in 2012, a poll of thirty-nine prominent US economists found that *not a single respondent* believed that on a gold standard, "price-stability and employment outcomes would be better for the average American" (Clark Center Forum 2012). Gold does not possess any intrinsic economic value; its economic value is just whatever the market is prepared to pay for it. And so far from being unusually stable in that value, research suggests that, if anything, the very perception that it is stable in times of crises increases its volatility in those situations (Corbet et al. 2020).

This points us, intriguingly, toward a skeptical solution to our conundrum. Following his mention of the elastic ruler, Wittgenstein remarks that its usefulness as a measure entirely depends on what exactly one is trying to do; "we could imagine a situation in which this was just what we wanted" (1953, I §5). This thought is taken up in commentary by Peter Hacker who, entirely coincidentally, notes that "[d]espite philosophers' qualms about the intelligibility of measuring with elastic rulers, we can readily understand it given an appropriate stage-setting. After all, we use elastic rulers ourselves all the time—in the fluctuating exchange rates of foreign currencies" (2015, 17).

I want, then, to take the metaphor of a "gold standard" newly seriously as an image for how a medical epistemology based on randomized control trials can work. It was wrongly thought that gold provided a stable and absolute value for currency, when in fact it did not; the value of gold is only ever the price that can be got for it at some particular place or time. We are accustomed to treating placebo-controlled RCTs as yielding stable and absolute measurement values, when in fact, again, they do not. Indeed, because what RCTs measure is intrinsically biological, we should not expect them to; the magnitudes of treatment effects, like everything else in biology, are fundamentally dynamic, and any particular RCT result is merely a snapshot, or a reification at one moment of a process in flux (see e.g., Dupré 2014; Dupré and Leonelli

2022). At most, like the price of gold, the result of a given trial should be regarded as *tolerably stable under favorable conditions*.

This means that our knowledge of medical efficacy and effectiveness is not best thought of by analogy with a catalogue of physical measurements, such as a road atlas, or the specification sheet for a racing bicycle. Rather, we should conceive of it in terms of something like a currency exchange, or another commodity or stock market. That is, clinical trial results form a floating network of contingent and dynamic comparative values. Some of these relations will be more stable than others at any given time, but they should never be mistaken for permanent or categorical states of affairs.

Given this analogy, it may be clarifying to briefly look at another neglected philosophical subfield here, the philosophy of economics and econometrics. The economy, as another sort of floating network of values, bears both similarities and dissimilarities as an object of study to the matrix of medical evidence. Economists have long sought to formulate or discover, not merely stable relationships between economic values, but underlying *laws* that govern them, reflecting some deeper economic reality. As Milton Friedman expressed this purportedly scientific attitude, economists seek “a way of looking at or interpreting or organizing the evidence that will reveal superficially disconnected and diverse phenomena to be manifestations of a more fundamental and relatively simple structure” (1953/1984, 231, quoted in Dupré 1993, 363).

As many critics of this optimistic view have observed, however, the issue it faces is, precisely, that “the evidence” is a complex and interrelated network of relations which, consequently, cannot be made to speak unambiguously with one voice (see e.g., Dupré 1993; Cartwright 1999; Boumans 2015; Skidelsky 2020). Controlled experiments cannot be performed on any significant scale, and because the network involves “in principle an infinite number of potential influences of which an (unknown) part is unknown and another part is not measurable, any representation will inevitably be incomplete, ‘inexact,’ however large and comprehensive the model will be” (Boumans 2015, 23). So if economics is to be in some sense scientific, it cannot be an “exact” science. For the Nobel Prize-winning econometrician Trygve Haavelmo, we might nevertheless “hope to find elements of invariance in economic life, upon which to establish permanent ‘laws,’” just insofar as under special, temporary (“*ceteris neglectis*”) conditions, the majority of potential influences are invariant and can be ignored (1944, 13, quoted in Boumans 2015, 106; see also Cartwright 1994, 72–73). Others frame the lesson more pessimistically; “[T]here are no ‘laws of economics’ valid at all times and places. At best, theories can lead to approximately reliable predictions over such time periods as other things stay the same” (Skidelsky 2020, 77–78). Or, more starkly again; “The most secure laws of economics are tendencies at best” (ibid., 2).

If the same or closely-related limitations apply to the network of medical trial data, their effect is considerably mitigated by an important difference; few if any medical researchers have understood their aim as being the discovery or elucidation of general predictive laws. While economists have often been accused of “physics envy,” medical research has been happy to operate on the stamp-collecting side of Rutherford’s famous dichotomy, cataloguing localized phenomena like responses to

therapeutic intervention without any particular theoretical ambitions to universality, or to the revelation of an underlying fundamental reality.

What I have argued in this article is, in effect, that we should not regard this catalogue as *accumulating* knowledge about those phenomena; new trial results are not to be straightforwardly added to the existing corpus. A more useful model, perhaps, would be that of a photo-album; we collect and collate our snapshots, and try to keep them up-to-date, but while we find the archive useful in observing patterns of change and stability over time, we do not expect the older images to add to our knowledge of how the people and places depicted currently stand. Each is merely a record; at some particular time and place, things stood thus-and-so. To conflate such reifications with the dynamic phenomena of interest, or to take them as revealing of fundamental structures, is not merely to make a category error; it is to risk potentially catastrophic failures of medical intervention by shooting at goalposts that have already moved (Dupré and Leonelli 2022).

To avoid such an outcome, several quite tentative and programmatic “best practice” recommendations can be made for the future conduct of medical trials, and publication of their results:⁸

First: Relational properties *are* relational, and this should be clearly communicated. That is, both scholarly and popular reporting on the findings of clinical trials should as far as possible eschew claims about absolute or categorical values (“New Treatment Halves Risk of Heart Disease”), and instead be explicit about their multivalently comparative nature, as well as about the particular comparators used.

Second: Expense permitting,⁹ multiarm and multisite trials are far preferable to single-site trials with only a placebo, no-treatment, or active control arm; they provide more data points, more points of triangulation, and more evidence of which relations are currently comparatively stable and which are not. To complicate matters, there is at least some evidence that the number of arms in a trial may influence the strength of any placebo effect (see Dworkin et al. 2005). While the authors of that study therefore recommend that trials use no more than three arms, I take the opposite view; all the more reason for all the more triangulation points!

Third: The exact control protocols used, especially regarding placebo arms, should be explicitly communicated in any scholarly or popular publication of trial results, in order that interexperimental comparability can be easily—if imprecisely—assessed. One might hope that this would form part of a stronger commitment to data-sharing among researchers more generally, but that is an argument for another occasion (see e.g., Staunton et al. 2021). The US National Institutes for Health’s *Placebo Response Drug Trials Survey* has been a notable step forward in this regard, though it was hampered by the reluctance of pharmaceutical companies to share what is often viewed as proprietary information.

⁸ These recommendations do, to at least some extent, reflect already-existing practice among researchers, even though that practice is not motivated by concerns about placebo response-variability. This should not be surprising, as we often have multiple independent reasons to do something; what the argument of this article provides, in such cases, is *additional and enhanced* reason to follow such procedures. I thank an anonymous reviewer for raising this point.

⁹ This is not at all a trivial qualification, though there is some recent evidence that the costs of medical trials have been systematically overestimated; see Bosely (2024).

Fourth: RCTs should be treated as having a “shelf-life”; although no hard cutoff point needs to be imposed, meta-analyses should assign a lower evidentiary weight the older a study is. Some do this already, but the practice should be standardized.

Throughout this discussion I have been principally concerned with trials of treatment *efficacy*—the strength of its effect under idealized trial conditions—rather than of its strength in “real-world” applications to the general public, or treatment *effectiveness*. While these exist on a continuum (e.g., Thorpe et al. 2009), effectiveness trials by their nature will often be observational in nature, such as case control or cohort studies; when they do take the form of RCTs, they will more often be active-controlled than placebo-controlled (Singal et al. 2014). Because observational studies don’t try to measure against placebo responses in the first place, the inherent inaccuracy of such measurements is not a problem for them. But where effectiveness trials are placebo- or active-controlled, the same concerns will apply as for efficacy trials. Comparative Effectiveness Research—defined as “the generation and synthesis of evidence that compares the benefits and harms of alternative methods to prevent, diagnose, treat, and monitor a clinical condition, or to improve the delivery of care” (Sox and Greenfield 2009)—has been a growing area of policy concern and academic research, and a significant attractor of funding; these guidelines have obvious relevance to that effort (see also Rothwell 2005).

I have made a skeptical argument here about the status of clinical trial results, as they are commonly understood. But I have also proposed a skeptical solution, in the tradition of Hume, Wittgenstein, and Kripke; we can keep the results, as long as that common understanding of their significance is changed. Doing so will not be without difficulty and expense, but it will leave us with a corpus of medical knowledge that is far more dynamic, far more responsive to the potential vagaries of treatment efficacy, and far more epistemically secure than what went before.

Acknowledgments. For Lorcan, who arrived during the writing process, and occasionally permitted it to continue; you’re the measure of my dreams.

I am grateful to an audience at the 2021 ISPHSSB meeting, hosted online by Cold Harbour Spring Laboratory, as well as to Elis Jones, Stephan Guttinger, Alex Broadbent, and Jim McCarron for helpful discussion of these issues. I’m also indebted to two anonymous reviewers for this journal for their feedback that I think significantly improved the article’s clarity.

This article is an outcome of the Research Project JOPS (Public Justification and Capability Pluralism), financed by the Croatian Science Foundation (HRZZ) (PI Elvio Baccarini; grant number: IP-2020-02-8073).

Funding Information. None to declare.

Declaration of Competing Interests. None to declare.

References

- Anderson, James A. 2006. “The Ethics and Science of Placebo-Controlled Trials: Assay Sensitivity and the Duhem-Quine Thesis.” *Journal of Medicine and Philosophy* 31:65–81. <https://doi.org/10.1080/03605310500499203>.
- Aulas, Jean-Jacques, and Itzhak A. Rosner. 2003. “Effets de la Prescription d’un Placebo ‘Annoncé’ [Efficacy of a Nonblind Placebo Prescription].” *L’Encéphale* 29:68–71.
- Bavbek, S., Ö. Aydın, Z.Ç. Sözen, and S. Yüksel. (2015), “Determinants of Nocebo Effect during Oral Drug Provocation Tests.” *Allergologia et Immunopathologia* 43 (4): 339–45. <https://doi.org/10.1016/j.aller.2014.04.008>.

- Beecher, Henry K. 1955. "The Powerful Placebo." *Journal of the American Medical Association* 159:1602–6. <http://doi.org/10.1001/jama.1955.02960340022006>.
- Bosely, Sarah. 2024. "Cost of Developing New Drugs May Be Far Lower Than Industry Claims, Trial Reveals." *The Guardian* April 25. <https://www.theguardian.com/global-development/2024/apr/25/cost-of-developing-new-drugs-may-be-far-lower-than-industry-claims-trial-reveals>.
- Boumans, Marcel. 2015. *Science Outside the Laboratory: Measurement in Field Science and Economics*. Oxford: Oxford University Press. <http://doi.org/10.1093/acprof:oso/9780199388288.001.0001>.
- Boussageon, Rémy, François Gueyffier, Theodora Bejan-Angoulvant, and Géraldine Felden-Dominiak. 2008. "Critique du Modèle Additif de l'Essai Clinique Randomisé." *Thérapies* 63 (1):29–35. <http://doi.org/10.2515/therapie.2008015>.
- Boussageon, Rémy, Jeremy Howick, Raphael Baron, Florian Naudet, Bruno Falissard, Ghina Harika-Germaneau, Issa Wassouf, François Gueyffier, Nemat Jaafari, and Clara Blanchard. 2022. "How Do They Add Up? The Interaction between the Placebo and Treatment Effect: A Systematic Review." *British Journal of Clinical Pharmacology* 88 (8):3638–56. <https://doi.org/10.1111/bcp.15345>.
- Campbell, Norman R. 1928. *An Account of the Principles of Measurements and Calculations*. London: Longmans Green.
- Campbell, Norman R. 1938. "Measurement and Its Importance for Philosophy." *Proceedings of the Aristotelian Society Supplementary Volume* 17 (1):121–51. <https://doi.org/10.1093/aristoteliansupp%2F17.1.121>.
- Cartwright, Nancy. 1994. *Nature's Capacities and Their Measurement*. Oxford: Oxford University Press. <http://doi.org/10.1093/0198235070.001.0001>.
- Cartwright, Nancy. 1999. *The Dappled World: A Study of the Boundaries of Science*. Cambridge: Cambridge University Press. <http://doi.org/10.1017/CBO9781139167093>.
- Cartwright, Nancy. 2007. "Are RCTs the Gold Standard?" *BioSocieties* 2:11–20. <http://doi.org/10.1017/S1745855207005029>.
- Chang, Hasok. 2004. *Inventing Temperature: Measurement and Scientific Progress*. Oxford: Oxford University Press. <http://doi.org/10.1093/0195171276.001.0001>.
- Clark Center Forum. 2012. "Surveys: Gold Standard, January 12 2012." Kent A. Clark Center for Global Markets website. <https://www.kentclarkcenter.org/surveys/gold-standard>.
- Corbet, Shaen, Charles Larkin, and Brian Lucey. 2020. "The Contagion Effects of the COVID-19 Pandemic: Evidence from Gold and Cryptocurrencies." *Finance Research Letters* 35:101554. <https://doi.org/10.1016/j.frl.2020.101554>.
- Deaton, Angus, and Nancy Cartwright. 2018. "Understanding and Misunderstanding Randomized Controlled Trials." *Social Science & Medicine* 210:2–21. <http://doi.org/10.1016/j.socscimed.2017.12.005>.
- Dupré, John. 1993. "Could There Be a Science of Economics?" *Midwest Studies in Philosophy* XVIII:363–78. <http://doi.org/10.1111/j.1475-4975.1993.tb00273.x>.
- Dupré, John. 2014. "A Process Ontology for Biology." *The Philosophers' Magazine* 67:81–88. <http://doi.org/10.5840/tpm201467117>.
- Dupré, John, and Sabina Leonelli. 2022. "Process Epistemology in the COVID-19 Era: Rethinking the Research Process to Avoid Dangerous Forms of Reification." *European Journal for Philosophy of Science* 12:20. <http://doi.org/10.1007/s13194-022-00450-4>.
- Dworkin, Robert H., Jennifer Katz, and Michael J. Gitlin. 2005. "Placebo Response in Clinical Trials of Depression and Its Implications for Research on Chronic Neuropathic Pain." *Neurology* 65 (Suppl. 4): S7e–S19. https://doi.org/10.1212/wnl.65.12_suppl.4.s7.
- Feinstein, Alvan R., and Ralph I. Horwitz. 1982. "Double Standards, Scientific Methods, and Epidemiologic Research." *New England Journal of Medicine* 307:1611–17. <https://doi.org/10.1056/nejm19821223072604>.
- Finkelstein, Ludwik. 2008. "Strength in Numbers." *Measurement + Control* 41 (1):21–24. <http://doi.org/10.1177/002029400804100105>.
- Finkelstein, Ludwik, and Mark S. Leaning. 1984. "A Review of the Fundamental Concepts of Measurement." *Measurement* 2 (1):24–34. [http://doi.org/10.1016/0263-2241\(84\)90020-4](http://doi.org/10.1016/0263-2241(84)90020-4).
- Freedman, Benjamin. 1987. "Equipoise and the Ethics of Clinical Research." *New England Journal of Medicine* 317 (3):141–45. <https://doi.org/10.1056/nejm198707163170304>.
- Friedman, Milton. 1953/1984. "The Methodology of Positive Economics." In *The Philosophy of Economics*, ed. Daniel M. Hausman. Cambridge: Cambridge University Press.

- Gotzsche, Peter C. 1995. "Concept of Placebo Should Be Discarded." *British Medical Journal* 311:1640–41. <https://doi.org/10.1136/bmj.311.7020.1640b>.
- Grünbaum, Adolf. 1986. "The Placebo Concept in Medicine and Psychiatry." *Psychological Medicine* 16: 19–38. <https://doi.org/10.1017/s0033291700002506>.
- Haavelmo, Trygve. 1944. *The Probability Approach in Econometrics*. Supplement to *Econometrica* 12 (July): S1–S115. <https://doi.org/10.2307/1906935>.
- Hacker, Peter M. S. 2015. "Forms of Life." *Nordic Wittgenstein Review* 4:1–20. <http://doi.org/10.15845/nwr.v4i0.3320>.
- Hall, Kathryn T., and Joseph Loscalzo. 2019. "Drug-Placebo Additivity in Randomized Clinical Trials." *Clinical Pharmacology & Therapeutics* 106 (6):1191–97. <https://doi.org/10.1002/cpt.1626>.
- Hill, Austin B. 1963. "Medical Ethics and Controlled Trials." *British Medical Journal* 1:1043–49. <https://doi.org/10.1136/bmj.1.5337.1043>.
- Ho, Dien. 2023. *What Placebos Teach Us about Health and Care: A Philosopher Pops a Pill*. Cambridge: Cambridge University Press. <http://doi.org/10.1017/9781009085496>.
- Hobsbawm, Eric. 1995. *The Age of Extremes: 1914–1991*. London: Abacus.
- Holland, Paul W. 1986. "Statistics and Causal Inference." *Journal of the American Statistical Association* 81:945–60. <http://doi.org/10.1080/01621459.1986.10478354>.
- Howick, Jeremy. 2017. "The Relativity of 'Placebos': Defending a Modified Version of Grünbaum's Definition." *Synthese* 194:1363–96. <https://doi.org/10.1007/s11229-015-1001-0>.
- Hume, David. 1748/1993. *An Enquiry Concerning Human Understanding*, ed. E. Steinberg. Indianapolis: Hackett Classics.
- ICH Expert Working Group. 2000. *ICH Harmonised Tripartite Guideline: Choice of Control Group and Related Issues in Clinical Trials*. International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use. https://database.ich.org/sites/default/files/E10_Guideline.pdf.
- Jones, David S., and Scott H. Podolsky. 2015. "The History and Fate of the Gold Standard." *The Lancet* 385:1502–3. [https://doi.org/10.1016/s0140-6736\(15\)60742-5](https://doi.org/10.1016/s0140-6736(15)60742-5).
- Krauss, Alexander. 2018. "Why All Randomised Controlled Trials Produce Biased Results." *Annals of Medicine* 50 (4):312–22. <https://doi.org/10.1080/07853890.2018.1453233>.
- Kripke, Saul. 1982. *Wittgenstein on Rules and Private Language*. Cambridge, MA: Harvard University Press.
- Kube, Tobias, and Winfried Rief. 2017. "Are Placebo and Drug-Specific Effects Additive? Questioning Basic Assumptions of Double-Blinded Randomized Clinical Trials and Presenting Novel Study Designs." *Drug Discovery Today* 22(4):729–35. <http://doi.org/10.1016/j.drudis.2016.11.022>.
- Louhiala, Pekka, Raimo Puustinen, and Harri Hemilä. 2013. "Impure Placebo as an Unsound Concept and Other Problems in the Paper by Howick et al." *PLoS ONE* 2013:e58247. <http://doi.org/10.13140/2.1.2591.8246>.
- Marx, Karl. 1993. *Grundrisse: Foundations of the Critique of Political Economy*. London: Penguin Classics.
- Meissner, Karin, Niko Kohls, and Luana Colloca. 2011. "Introduction to Placebo Effects in Medicine: Mechanisms and Clinical Implications." *Philosophical Transactions of the Royal Society B* 366:1783–89. <https://doi.org/10.1098/rstb.2010.0414>.
- Miller, Franklin G., and Howard Brody. 2003. "A Critique of Clinical Equipoise: Therapeutic Misconception in the Ethics of Clinical Trials." *The Hastings Center Report* 33 (1):19–28. <https://doi.org/10.2307/3528434>.
- Modell, W., and Margaret Garrett. 1960. "Interactions between Pharmacodynamic and Placebo Effects in Drug Evaluations in Man." *Nature* 185 (4712):538–39. <https://doi.org/10.1038/185538a0>.
- Moerman, Daniel E. 2000. "Cultural Variations in the Placebo Effect: Ulcers, Anxiety, and Blood Pressure." *Medical Anthropology Quarterly* 14 (1):51–72. <https://doi.org/10.1525/maq.2000.14.1.51>.
- Okusogu, Chika, Yang Wang, Titilola Akintola, Nathaniel R. Haycock, Nandini Raghuraman, Joel D. Greenspan, Jane Phillips, Susan G. Dorsey, Claudia M. Campbell, and Luana Colloca. 2020. "Placebo Hypoalgesia: Racial Differences." *Pain* 161:1872–83. <https://doi.org/10.1097/j.pain.0000000000001876>.
- Olson, Elizabeth M., Titilola Akintola, Jane Phillips, Maxie Blasini, Nathaniel R. Haycock, Pedro E. Martinez, Joel D. Greenspan, Susan G. Dorsey, Yang Wang, and Luana Colloca. 2021. "Effects of Sex on Placebo Effects in Chronic Pain Participants: A Cross-Sectional Study." *Pain* 162:531–42. <https://psycnet.apa.org/doi/10.1097/j.pain.0000000000002038>.

- Olson, Jay A., Michael Lifshitz, Amir Raz, and Samuel P. L. Veissière. 2021. "Super Placebos: A Feasibility Study Combining Contextual Factors to Promote Placebo Effects." *Frontiers in Psychiatry* 12:644825. <https://doi.org/10.3389/fpsy.2021.644825>.
- Park, Lee Crandall, and Lino Covi. 1965. "Nonblind Placebo Trial." *Archives of General Psychiatry* 12 (4): 336–45. <https://doi.org/10.1001/archpsyc.1965.01720340008002>.
- Paul, L. A., and Kieran Healy. 2018. "Transformative Treatments." *Nous* 52 (2): 320–35. <https://doi.org/10.1111/nous.12180>.
- Quitkin, Frederic M. 1999. "Placebos, Drug Effects, and Study Design: A Clinician's Guide." *The American Journal of Psychiatry* 156 (6):829–36. <https://doi.org/10.1176/ajp.156.6.829>.
- Reichenbach, Hans. 1958. *The Philosophy of Space and Time*. New York: Courier Dover.
- Rhees, Rush. 1970. "On Continuity: Wittgenstein's Ideas, 1938." In *Discussions of Wittgenstein*, 104–57. London: Routledge.
- Riordan, Sally. 2015. "The Objectivity of Scientific Measures." *Studies in History and Philosophy of Science* 50:38–47. <https://doi.org/10.1016/j.shpsa.2014.09.005>.
- Rothman, Kenneth J., and Karin B. Michels. 1994. "The Continuing Unethical Use of Placebo Controls." *New England Journal of Medicine* 331:394–98. <https://doi.org/10.1056/nejm199408113310611>.
- Rothwell, Peter M. 2005. "External Validity of Randomised Controlled Trials: 'To Whom do the Results of this Trial Apply?'" *The Lancet* 365 (9453):82–93. [https://doi.org/10.1016/s0140-6736\(04\)17670-8](https://doi.org/10.1016/s0140-6736(04)17670-8).
- Rubin, Donald B. 1974. "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies." *Journal of Educational Psychology* 66:688–701. <http://doi.org/10.1037/h0037350>.
- Rubin, Donald B. 2005. "Causal Inference Using Potential Outcomes." *Journal of the American Statistical Association* 100:322–31. <https://doi.org/10.1198/016214504000001880>.
- Sandra, Dasha A., Jay A. Olson, Ellen J. Langer, and Mathieu Roy. 2023. "Presenting a Sham Treatment as Personalised Increases the Placebo Effect in a Randomised Controlled Trial." *eLife* 12:e84691. <https://doi.org/10.7554/eLife.84691>.
- Schroeder, Severin. 2015. "Mathematics and Forms of Life." *Nordic Wittgenstein Review* 4:111–30. <http://doi.org/10.15845/nwr.v4i0.3357>.
- Shapiro, Arthur K. 1968. "Semantics of the Placebo." *Psychiatric Quarterly* 42:653–96. <https://doi.org/10.1007/BF01564309>.
- Singal, Amit G., Peter D. R. Higgins, and Akbar K. Waljee. 2014. "A Primer on Effectiveness and Efficacy Trials." *Clinical and Translational Gastroenterology* 5(1):e45. <https://doi.org/10.1038/ctg.2013.13>.
- Skidelsky, Robert. 2020. *What's Wrong with Economics? A Primer for the Perplexed*. New Haven: Yale University Press. <https://doi.org/10.12987/9780300252767>.
- Smith, Ben. 2019. *Doggerland*. London: 4th Estate.
- Sox, Harold C., and Sheldon Greenfield. 2009. "Comparative Effectiveness Research: A Report from the Institute of Medicine." *Annals of Internal Medicine* 151 (3):203–5. <https://doi.org/10.7326/0003-4819-151-3-200908040-00125>.
- Staunton, Cíara, Carlos Andrés Barragán, Stefano Canali, Calvin Ho, Sabina Leonelli, Matthew Mayernik, Barbara Prainsack, and Ambroise Wonkham. 2021. "Open Science, Data Sharing and Solidarity: Who Benefits?" *History and Philosophy of the Life Sciences* 43 (4):115. <https://doi.org/10.1007/s40656-021-00468-6>.
- Stevens, Stanley S. 1946. "On the Theory of Scales of Measurement." *Science* 103:677–80. <https://doi.org/10.1126/science.103.2684.677>.
- Tal, Eran. 2013. "Old and New Problems in Philosophy of Measurement." *Philosophy Compass* 8 (12): 1159–73. <https://doi.org/10.1111/phc3.12089>.
- Temple, Robert, and Susan S. Ellenberg. 2000. "Placebo-Controlled Trials and Active-Control Trials in the Evaluation of New Treatments, Part I: Ethical and Scientific Issues." *Annals of Internal Medicine* 133 (6):455–63. <https://doi.org/10.7326/0003-4819-133-6-200009190-00014>.
- Thomson, William. 1889. "Electrical Units of Measurement." In *Popular Lectures and Addresses, Vol. 1: Constitution of Matter*, 73–136. London: Macmillan.
- Thorpe, Kevin E., Merrick Zwarenstein, Andrew D. Oxman, Shaun Treweek, Curt D. Furberg, Douglas G. Altman, Seán Tunis, Eduardo Bergel, Ian Harvey, David J. Magid, and Kalipso Chalkidou. 2009. "A Pragmatic-Explanatory Continuum Indicator Summary (PRECIS): A Tool to Help Trial Designers." *Journal of Clinical Epidemiology* 62 (5):464–75. <https://doi.org/10.1016/j.jclinepi.2008.12.011>.

- Tuttle, Alexander H., Sarasa Tohyama, Tim Ramsay, Jonathan Kimmelman, Petra Schweinhardt, Gary J. Bennett, and Jeffrey S. Mogil. 2015. "Increasing Placebo Responses over Time in U.S. Clinical Trials of Neuropathic Pain." *Pain* 156 (12):2616–26. <https://doi.org/10.1097/j.pain.0000000000000333>.
- Weimer, Katja, Luana Colloca, and Paul Enck, P. 2015. "Age and Sex as Moderators of the Placebo Response —An Evaluation of Systematic Reviews and Meta-Analyses across Medicine." *Gerontology* 61 (2):97–108. <https://doi.org/10.1159/000365248>.
- Wittgenstein, Ludwig. 1953. *Philosophical Investigations*, trans. G. E. M. Anscombe. London: Basil Blackwell.
- Wittgenstein, Ludwig. 1978. *Remarks on the Foundations of Mathematics*, 3rd ed., eds. G. E. M. Anscombe, Rush Rhees, and G. H. von Wright. London: Basil Blackwell.
- World Medical Association. 2013. "WMA Declaration of Helsinki: Ethical Principles for Medical Research Involving Human Subjects." *World Medical Association* <https://www.wma.net/policies-post/wma-declaration-of-helsinki-ethical-principles-for-medical-research-involving-human-subjects/>.
- Wright, Crispin. 1980. *Wittgenstein on the Foundations of Mathematics*. Cambridge, MA: Harvard University Press.