Measuring Effectiveness

Jacob Stegenga

Final draft. June 9, 2015
Studies in History and Philosophy of Biological and Biomedical Sciences 54: 62-71

Abstract

Measuring the effectiveness of medical interventions faces three epistemological challenges: the choice of good measuring instruments, the use of appropriate analytic measures, and the use of a reliable method of extrapolating measures from an experimental context to a more general context. In practice each of these challenges contributes to overestimating the effectiveness of medical interventions. These challenges suggest the need for corrective normative principles. The instruments employed in clinical research should measure patient-relevant and disease-specific parameters, and should not be sensitive to parameters that are only indirectly relevant. Effectiveness always should be measured and reported in absolute terms (using measures such as 'absolute risk reduction'), and only sometimes should effectiveness also be measured and reported in relative terms (using measures such as 'relative risk reduction')—employment of relative measures promotes an informal fallacy akin to the base-rate fallacy, which can be exploited to exaggerate claims of effectiveness. Finally, extrapolating from research settings to clinical settings should more rigorously take into account possible ways in which the intervention in question can fail to be effective in a target population.

Acknowledgements

For extensive commentary on drafts I am grateful to Alex Broadbent, Luis Flores, Jonathan Fuller, Jonah Schupbach, and Eran Tal. I am also grateful for discussion with audiences at the 2014 Philosophy of Science Association meeting, the 2014 London workshop on Prediction in Epidemiology, the 2014 Helsinki workshop on Evidence in Science and Epistemology, and the University of Victoria Philosophy Colloquium.

1 Introduction

Much clinical research is designed to estimate the effectiveness of medical interventions. The details of this measurement procedure are interesting in their own right, and are perhaps more nuanced and complicated than many suppose. I describe some of these details in what follows, and argue that there are three widespread problems in measuring the effectiveness of medical interventions: the use of poor measuring instruments, the use of misleading analytic measures, and the assumption that measurements in an experimental setting are sufficient to infer properties of a general capacity of effectiveness. Each of these problems contributes to overestimating the effectiveness of medical interventions. The problems naturally suggest the need for corrective normative principles—medical research should use appropriate measuring instruments, truth-conducive analytic measures, and reliable methods of extrapolation. The employment of such principles would generally lead to lower—yet more accurate—estimates of the effectiveness of medical interventions than is presently the case.

By far the most common method for measuring effectiveness of medical interventions is the clinical trial.[1] A standard clinical trial involves administering a potential medical intervention at a particular dose to one group of subjects (the experimental group), administering a placebo or competitor intervention to another group of subjects (the control group), measuring one or more parameters of the subjects, comparing the values of those parameters between the two groups, and inferring a general effectiveness capacity from the difference in values of the parameters between the groups. Clinical trials usually have methodological safeguards to minimize systematic error, most prominently including the random allocation of subjects to groups, and concealment of the group assignment from both the investigators and the subjects. But these methodological details aside, the measurement of effectiveness itself involves three steps: the use of a measuring instrument (or a measuring technique more generally), the analysis of measured values, and the extrapolation of analyzed values to a target population.

Effectiveness of medical interventions is a causal capacity to modify properties of patients. This is not an intrinsic causal capacity; effectiveness is a relational property in which the relata are a causal capacity of the intervention and properties of a defined class of people. The properties that must be modulated by a medical intervention in order for that intervention to be deemed effective are either the constitutive causal basis of a disease or symptoms of a disease that cause harm to those with that disease. I defend this in the companion article to this one ('Effectiveness of Medical Interventions', published in this

[1] As I argue below, the exclusion of evidence from other kinds of methods in the measurement of effectiveness is a significant epistemic limitation. But since this reliance on clinical trials (and only clinical trials) is so ubiquitous, I maintain, for now, a narrow focus on this method.

issue), in which I call these two individually sufficient conditions for effectiveness *CAUSAL TARGET OF EFFECTIVENESS* and *NORMATIVE TARGET OF EFFECTIVENESS*. In the companion article my aim is to articulate a defensible view of what effectiveness as a measurand is (a conceptual and metaphysical question), whereas in the present article my aim is to articulate limitations on how we measure that measurand—a distinct epistemic question. In the companion article I rely on the idealization that effectiveness is a binary notion; this allowed me to explore facets of effectiveness without undue complications. But of course, effectiveness is a property to measured.

For any measurement one needs a measuring instrument. In clinical practice and clinical research various kinds of instruments are employed to measure various kinds of parameters, including subjective patient-reported parameters (such as reports of well-being), physician-reported parameters (such as appearance of lethargy), institutional parameters (such as number of days in an intensive care unit), and physiological parameters (such as blood sugar concentrations). For example, the Hamilton Rating Scale for Depression (discussed in further detail below) measures several of these kinds of outcomes, including a patient's report of sadness and quality of sleep, a physician's assessment of the patient's fidgetiness, and physiological correlates of anxiety. Sometimes the outcome of interest in a clinical trial is simple, like an event such as death, in which case the appropriate measuring instrument is whatever is required to determine that the event has occurred. I will use the term 'instrument' very broadly to include any tool or technique employed to estimate values of measurands. In §2 I describe various examples of measuring instruments, and argue that many such instruments in clinical research are not very good, because at best they measure proxies of the parameter of interest, or at worst are irrelevant to the parameter of interest.

Once parameters are chosen and instruments have been employed to assign values to those parameters among subjects in a clinical trial, those values must be interpreted in some way to assess whether, and if so to what extent, an intervention modifies the values of those parameters. Several descriptive statistics are widely employed in medical science as measures of effectiveness; these are called 'outcome measures', while the numerical outputs of outcome measures are often called 'effect sizes'. In §3 I describe several basic outcome measures and argue that the most widely employed class of outcome measures is misleading. From the perspective of a patient or a physician who is deciding whether or not to use or prescribe a particular treatment, the best outcome measures are so-called 'absolute' measures, or 'difference' measures, which, unlike 'relative' or 'ratio' measures, take into account the baseline value of whatever parameter is being measured.

The aim of measuring the effectiveness of medical interventions is to aid in decisions regarding treatment, which involves predicting outcomes in target patient populations (§4). One method for making such predictions is simple extrapolation from the quantitative results of clinical trials to a target population. Simple extrapolation is often

implicitly employed in medical decision-making, and is sometimes explicitly defended as a reliable method for extrapolation. But I argue that simple extrapolation is unreliable, and it tends to overestimate the effectiveness of medical interventions in target populations.

Thus, clinical research involves a chain of measurands, in which the value of one measurand is used to infer the value of the next measurand in the chain. This is not a unique scenario for the epistemology of measurement—measuring the temperature in my backyard involves measuring the height of mercury in a glass tube; measuring the rate of expansion of the universe involves measuring Hubble's Constant by measuring wavelengths of light undergoing redshift.[2] The ultimate measurand of interest in clinical research is the effectiveness of a medical intervention. Estimating this measurand is based (at least in part) on a prior measurand: the capacity of the medical intervention, in a controlled experimental setting, to cause a difference in the value of the parameter of interest between the experimental group and the control group. This in turn involves measurement of the value of that very parameter in those subjects. At each of the three links of this chain of measurands there are methodological challenges that occupy the attention of clinical scientists and are often not adequately resolved in clinical research.

In short, the measurement of effectiveness of medical interventions faces three methodological challenges, associated with the choice of measuring instrument (§2), outcome measure (§3), and method of extrapolation (§4). I am not the first to note these challenges. But in what follows I argue that in practice these methodological challenges contribute to overestimating the effectiveness of medical interventions. If these challenges were better addressed, estimates of the effectiveness of medical interventions would be more accurate, and lower than they are now.

2 Instruments

To determine the values of parameters of subjects in the experimental and control groups of a clinical trial, one needs a measuring instrument. Such instruments can vary in a number of important respects. These instruments can be simple, particularly when the measurand is an event (such as death), or they can be multifaceted, particularly when the measurand is characterized by medical constructs (such as depression). Another dimension of these instruments is their inferential directness: some instruments involve relatively direct measures of the measurands of interest, in that the value determined by the instrument requires only a few (usually reliable) inferences to determine the value of the measurand of interest. Other instruments are inferentially indirect, in that they are

[2] For recent work on the epistemology of measurement, see (Alexandrova, 2008), (Tal, 2011), (Tal, forthcoming), (Teller, 2013), and (van Fraassen, 2008).

measures of a proxy of the measurand of interest, and the measurement procedure requires more inferences (which are often less reliable) from the value of the measured parameter to the value of the measurand of interest. In the clinical literature such proxy parameters are called 'surrogate outcomes'. As with all measuring instruments, two central desiderata are sensitivity and specificity: a measuring instrument should be sensitive to the true values of the measurand of interest, and should be sensitive only to such values. The employment of certain instruments, some of which are widely used in clinical research, contributes to frequent overestimations of the effectiveness of experimental medical interventions.

Here is an example of a relatively indirect instrument which, it turns out, is nonspecific to values of the measurand of interest. Some evidence suggests that certain interventions can reduce the 'white lesions' that are said to be physiological markers ('biomarkers') of multiple sclerosis (MS). White lesions are the result of the demyelination of the sheaths that surround the axons of neurons, and are not themselves the ultimate causes of MS (which remains unknown). The hope, however, is that if white lesions are a proximal cause of MS symptoms (below I note that this is doubtful, since white lesions are an *effect* of the demyelination of axon sheaths, and so could be just another symptom of MS, albeit at a cellular level), then mitigating white lesions will mitigate MS symptoms. Some trials on novel interventions for MS have employed, as a measured outcome, the amount of white lesions in subjects, and have showed that some novel interventions have the capacity to decrease the amount of white lesions. Since the ultimate goal is the mitigation of MS symptoms, the instruments in such trials measure a proxy of the measurand of interest: white lesions as a surrogate outcome of patient-level MS symptoms (see (Lavery, Verhey, & Waldman, 2014) for a discussion of various outcomes measured in trials of interventions for MS). Under the assumption that mitigating white lesions will mitigate patient-level MS symptoms, the results of this measurement procedure license an inference that the drugs under investigation are effective at mitigating MS symptoms. Unfortunately, the available evidence suggests that such drugs have little impact on MS symptoms. Thus, the use of white lesions as a proxy for patient-level MS symptoms is nonspecific, because it is sensitive to values of parameters that are only weakly correlated with the measurand of interest. The inferential assumption noted above is probably false: demyelination of axon sheaths is probably best construed as a common cause of white lesions and patient-level MS symptoms, and since the direction of causal relevance is not (as far as we can tell) from white lesions to patient-level MS symptoms, measuring a reduction in white lesions does not warrant an inference about the reduction of patient-level MS symptoms. The measuring instrument in this case overestimates the true value of the measurand of interest.

Here is an example of a multifaceted instrument. The Hamilton Depression Rating Scale (HAMD), one of the most commonly employed instruments in clinical trials testing

the effectiveness of antidepressants, is a multifaceted instrument. It is a questionnaire composed of 17 questions, each of which has between three to five possible answers with a corresponding numerical score, which taken together are thought by some to measure the severity of depression (there are an additional four questions which do not contribute to the score).[3] A total score can range from 0 to 52 points, and scores are interpreted in terms of severity of depression as follows: 0 - 7: normal; 8 - 13: mild; 14 - 18: moderate; 19 - 22: severe; ≥ 23: very severe. HAMD scores are determined for subjects in a clinical trial, and if an antidepressant is effective one ought to observe a greater decrease in HAMD score for subjects in the experimental group of the trial compared with subjects in the control group.[4] However, the HAMD is an nonspecific instrument with regard to the measurand of interest, namely, intensity of depression. That is because many of the questions included on the HAMD are largely irrelevant to this measurand.

Some of these questions probe core elements of depression, albeit at a coarse grain. For instance, the question on 'suicidality' is scored as follows: "0 = Absent. 1 = Feels life is not worth living. 2 = Wishes he were dead or any thoughts of possible death to self. 3 = Suicidal ideas or gesture. 4 = Attempts at suicide (any serious attempt rates 4)." Thus the greater the degree of suicidality of a subject, the greater the HAMD score.[5] Other questions in the HAMD probe features of a person's state that are less central to depression. For instance, there are three questions regarding insomnia, corresponding to three phases of night (early, mid, and late), and there are a possible six points available for these questions. Similarly, there are four possible points associated with fidgeting. Thus, if a novel alleged intervention for depression causes people to sleep better and fidget less, the corresponding HAMD reduction could be ten points (to put this in perspective, some clinical guidelines have held that a reduction of three points on the HAMD scale by an experimental intervention entails that the intervention is effective). A small improvement in sleep or a decrease in fidgeting caused by an experimental drug would warrant approval of this drug as an effective antidepressant (despite the fact that the drug might

[3] There are various versions of the HAMD, but here I describe the original proposed by (Hamilton, 1960). This instrument has been hugely influential in clinical psychiatric research. Note that sometimes in the clinical literature alternative abbreviations are used for the scale, including HDRS and HRSD.

[4] (van Fraassen, 2008) argues that when measuring physical property x the questions 'what is x?' and 'what is a good measurement of x?' are fundamentally dependent on one another. (McClimans, 2013) argues that this is similarly true for the measurement of psychological phenomena, and notes the importance of this for the validity of outcome measures based on patient reports. The example of the HAMD scale for measuring severity of depression exemplifies this problem of measurement circularity. For a now canonical statement of the problem, see (Chang, 2004).

[5] However, the question is insensitive to relatively important differences in the degree of one's possible suicidality; for instance, the difference between 0 points and 1 point is stipulated as the difference between a feeling of no suicidality at all to an unqualified feeling that life is not worth living—a phenomenological leap.

not mitigate any of the fundamental symptoms of depression, such as low mood, anhedonia, and feelings of worthlessness, guilt, and hopelessness). The HAMD is an example of a multifaceted instrument which could overestimate effectiveness of interventions due to the instrument not being sufficiently specific.[6]

A curious HAMD question is titled 'Insight', and is scored as follows: "0 = Acknowledges being depressed and ill. 1 = Acknowledges illness but attributes causes to bad food, climate, overwork, virus, need for rest, etc. 2 = Denies being ill at all." Thus a tired patient who suspects that her illness is caused by a gluten allergy automatically gets an extra point on her HAMD score. Philosophical critics and rugged cowboys, beware! Denying one's alleged depression earns you *two* extra points. A more prosaic way to state this measurement problem: if an experimental intervention causes a subject who initially denies being depressed to then claim that they are depressed—say, because the intervention itself causes symptoms of depression—then that subject's HAMD score would go *down* by two points, thereby making the intervention appear to be an effective antidepressant, despite itself causing the symptoms of depression.

Instruments for measuring the effectiveness of medical interventions should be sensitive to a time-index relative to the characteristics of the particular disease being treated (the term of art in medicine is the 'clinical course' of the disease). Unfortunately such instruments are often insensitive to the clinical course of a disease. For instance, when researchers were testing high-dose chemotherapy for breast cancer, they assessed presence of cancer after 18 months of treatment (see (Brownlee, 2008) for a discussion of this episode). This temporal range was adopted from blood cancers, in which high-dose chemotherapy is effective. After 18 months it appeared that the high-dose chemotherapy had prevented recurrence of breast cancers. However, breast cancers grow slower than blood cancers, and so 18 months was an inappropriately short time to measure the outcome of the therapy. Later studies that used a longer temporal range found that high-dose chemotherapy did more harm than good for breast cancers. The physiological difference in growth rates between cancer types explains why high-dose chemotherapy is more effective in blood cancers than in breast cancers: since chemotherapeutic drugs operate by interfering with mechanisms of cell division, cells that divide rapidly are more susceptible to chemotherapy (slower growing tumors are less susceptible to chemotherapy). The initial studies that suggested that high-dose chemotherapy is effective for breast cancer employed an instrument that was not sufficiently sensitive to the temporality of the disease.

---

[6] One might hold that among patients with depression, their insomnia is likely to be caused by their disease, and thus if a drug causes a patient to sleep better then either the drug is (i) intervening in the pathophysiology of depression, or at least is (ii) offering relief of symptoms of depression. (i) strikes me as excessively optimistic, but (ii) is surely reasonable. Such a drug then should be considered a soporific rather than an 'antidepressant'.

Another example in which an inappropriate temporal range has been used to measure effectiveness are trials testing the effectiveness of methylphenidate (Ritalin) to treat attention deficit hyperactivity disorder (ADHD). The vast majority of such trials have only lasted a few weeks, and meta-analyses of these trials suggest that methylphenidate has a small but positive effect on ADHD symptoms in the short term (see, e.g. (Schachter, Pham, King, Langford, & Moher, 2001)). But in studies that follow-up on patients from the longest trial performed thus far—the 14-month MTA trial of The National Institute of Mental Health (NIMH)—there is no beneficial difference in ADHD symptoms among children who had been on methylphenidate compared to children who had not. There are, however, apparent harms caused by methylphenidate in the long run, including a decrease in body height and mass. These follow-ups have been done at three years and eight years after the trial ((Jensen et al., 2007), (Molina et al., 2009)). Thus an inappropriately short temporal range of most studies of the effectiveness of methylphenidate contribute to an overestimation of the effectiveness of the drug (and an underestimation of its harm profile).

The use of indirect instruments also contributes to overestimation of effectiveness, since the causal link between the measurand used in indirect instruments and the measurand of interest is often not as tight as one might hope. For example, high cholesterol levels are thought by many to be a cause of heart disease, and so to avoid heart disease, cholesterol-lowering drugs are prescribed to many people. Clinical trials have shown that these drugs are effective at lowering cholesterol. With respect to the end of avoiding heart disease, these clinical trials employ an indirect instrument (measurement of cholesterol levels), and under the assumption that high cholesterol levels cause heart disease, the evidence from these trials warrants an inference that cholesterol-lowering drugs are effective at mitigating heart disease. Unfortunately, trials that employ the more direct instrument of measuring heart disease have shown that these drugs are barely effective at mitigating heart disease (see (Moynihan & Cassels, 2005)). To use the analysis from the companion article (Stegenga, forthcoming), the problem with using indirect instruments to measure surrogate outcomes is that an intervention that modulates a surrogate outcome might not satisfy *CAUSAL TARGET OF EFFECTIVENESS*, because the surrogate outcome is not the constitutive causal basis of the disease nor is likely to be causally prior to the constitutive causal basis of the disease, and similarly, an intervention that modulates a surrogate outcome might not satisfy *NORMATIVE TARGET OF EFFECTIVENESS*, because by definition surrogates are stand-ins for the patient-level outcomes that matter (the normative basis of the disease).

McClimans (2010) notes that there are thousands of such measuring instruments in medicine, and yet the theoretical underpinning of such instruments is often poorly understood, and clinical researchers often ignore the consequences of this. I have argued here that problems of measuring instruments in clinical research contribute to an

overestimation of the effectiveness of medical interventions. In principle, the measurement bias introduced by instruments in clinical research is symmetric with respect to estimating effectiveness: the above problems could lead to underestimation of effectiveness just as often as they lead to overestimation of effectiveness. For example, just as the temporal range of a study can be insufficiently short to observe ineffectiveness in the longer term, the temporal range of a study can be too short to observe effectiveness in the longer term. Similarly, clinical studies on experimental antidepressants could employ an instrument that is relatively insensitive to changes in a person's core symptoms of depression, in which case such studies would tend to underestimate the effectiveness of the tested antidepressants. Trial designers, I assume, are well aware of the methodological details regarding their measuring instruments which generate a trade-off between the predilection of a trial to overestimate effectiveness and the predilection of a trial to underestimate effectiveness. Moreover, trial designers have strong motive to err on the side of overestimating rather than underestimating effectiveness—this is a contingent sociological point based on pressure to publish among academic scientists and pressure to develop profitable products among corporate scientists. Since trial designers have strong motive to err on the side of overestimating effectiveness, they tend to do so, as illustrated by the above examples.

Once evidence is gathered by the use of such measuring instruments, the evidence is often analyzed and presented in such a way as to make the experimental interventions appear more effective than they are—a problem I now turn to.

3 Measures

Many 'outcome measures' are employed in clinical research. An outcome measure is an abstract formal statement describing a relation between the value of the measurand in the control group and the value of the measurand in the experimental group. When particular substantive values for such measurands are substituted into an outcome measure, the result is a quantitative estimation of the strength of an alleged causal relation—this quantity is usually called an 'effect size'.[7] There are outcome measures for both continuous and dichotomous parameters.

If the measured parameters are continuous (such as blood sugar concentration), a common outcome measure is the standardized mean difference (SMD):

$$SMD = (\mu_1 - \mu_2)/\sigma$$

[7] How an effect size relates to the strength of a causal relation is a tricky problem beyond the scope of the present article. See (Broadbent, 2013) for a discussion of what he calls, in the epidemiological context, 'puzzles of attributability'.

where μ_1 is the mean value of a parameter of interest for the experimental group, μ_2 is the mean value of the same parameter for the control group, and σ is a measure of the variance of the value of the parameter (for some statistics σ is measured in the control group and for other statistics the variance from both groups is pooled to determine σ). A ubiquitous practice in medical research is to use SMD as the basis of more complicated analytic statistics, such as the t-test, in the service of null hypothesis testing.[8] The simple measure $\mu_1 - \mu_2$ is important because it measures the absolute difference between mean values of the parameter of interest.

For both continuous and dichotomous parameters, the choice of outcome measure is important and can have significant influence on the estimation of effectiveness. The basic issue I discuss below is salient for both kinds of parameters. However, the point can be made more simply by focusing on dichotomous parameters (similarly, for simplicity I ignore discrete non-binary parameters).

If the measured parameters are dichotomous (such as death), standard outcome measures include the odds ratio, relative risk (sometimes called risk ratio), relative risk reduction, risk difference (sometimes called absolute risk reduction), and number needed to treat. To define these, one constructs a two-by-two table for a study that has an experimental group (E) composed of subjects who receive the experimental intervention, and a control group (C) composed of subjects who do not receive the experimental intervention (perhaps they receive a placebo), in which a binary outcome is measured as present (Y) or absent (N), where the number of subjects with each outcome in each group is represented by letters (a-d), as follows:

Group	Outcome	
	Y	N
E	a	b
C	c	d

Relative risk (RR) is defined as:

$$RR = [a/(a+b)] / [c/(c+d)]$$

Relative risk reduction (RRR) is defined as:

$$RRR = [[a/(a+b)] - [c/(c+d)]] / [c/(c+d)]$$

Risk difference (RD) is defined as:

$$RD = a/(a+b) - c/(c+d)$$

Number needed to treat (NNT) is defined as:

$$NNT = 1 / [[a/(a+b)] - [c/(c+d)]]$$

[8] A problem with null hypothesis testing pertinent to the present discussion is that a null hypothesis test only reports the probability that the observed difference between parameter means in the two populations was due to chance, and does not provide any added information about the *degree* of effectiveness.

It also can be useful to define these in terms of conditional probabilities. The probability of a subject having a Y outcome given that the subject is in group E, P(Y|E), is a/(a+b), and likewise, the probability of having a Y outcome given that the subject is in group C, P(Y|C), is c/(c+d). Thus, for example, we have:

RR = P(Y|E)/P(Y|C)

RD = P(Y|E) - P(Y|C)

A widespread and misguided practice is to report RR or RRR but not RD or NNT. The over-reliance on relative outcome measures in epidemiology is dubbed 'risk relativism' by Broadbent (2013). Broadbent canvasses several alleged justifications for the widespread use of relative measures like RR and finds them all wanting. He also notes that some epidemiologists have begun to urge more frequent use of absolute measures. My concern here is not with alleged justifications for risk relativism, but rather with a nefarious consequence of risk relativism. Employment of relative measures, such as RR or RRR, promotes the base-rate fallacy (Worrall, 2010). Both physicians and patients overestimate the effectiveness of medical interventions when presented with only relative measures, and their estimates are more accurate when they are presented with both relative and absolute measures or with absolute measures alone. This finding has been replicated many times in different contexts.[9]

To see that relative measures promote the base-rate fallacy, consider the following. Suppose Y is the beneficial outcome in question. The question of central concern for a patient is: to what extent would using this particular intervention change the probability of me having the outcome in question? Two epistemological notions are pertinent here: *change* and *probability*. A faithful way to represent this is to multiply two factors: a factor that represents the difference-making capacity of the intervention, and a factor that represents the baseline probability of the beneficial outcome in question.

By applying Bayes' Theorem, RR is equivalent to:

RR = [P(E|Y)P(Y)/P(E)] / [P(C|Y)P(Y)/P(C)]

= [P(E|Y)/P(E)] / [P(C|Y)/P(C)]

The baseline probability of having outcome Y, P(Y), has fallen out of the equation. Thus RR is not sensitive to P(Y).

In contrast, consider RD. By applying Bayes' Theorem, RD is equivalent to:

RD = [P(E|Y)P(Y)/P(E)] - [P(C|Y)P(Y)/P(C)]

= P(Y)[[P(E|Y)/P(E)] - [P(C|Y)/P(C)]]

The leftmost multiplicand just is the prior probability of Y. Thus RD is sensitive to P(Y). The rightmost multiplicand is a representation of the extent to which consuming the

[9] See, as examples: (Nexøe, Gyrd-Hansen, Kragstrup, Kristiansen, & Nielsen, 2002), (Forrow, Taylor, & Arnold, 1992), (Naylor, Chen, & Strauss, 1992), (Sorensen, Gyrd-Hansen, Kristiansen, Nexoe, & Nielsen, 2008), and (Bobbio, Demichelis, & Giustetto, 1994).

intervention changes the probability of Y.[10] One can see this perhaps more clearly by applying Bayes' Theorem once again, to the rightmost multiplicand:

$$RD = P(Y)[[P(Y|E)/P(Y)] - [P(Y|C)/P(Y)]]$$

The terms in the rightmost multiplicand are intuitive representations of the difference making capacity of the experimental intervention and control intervention, respectively. Of course, these quantities are derived from the hypothetical study in question, and in §4 I argue that care should be applied when extrapolating from the results of a study to making an inference about how beneficial an intervention will be for a particular patient. Nevertheless, as suggested above, to address the patient's central question articulated above, we should have a measure that represents the capacity of an intervention to change the probability of the beneficial outcome in question. RR does not do this. RD does.

Here is a related, decision-theoretic argument in favor of RD over RR. (Worrall, 2010) rightly notes that the choice of using a medical intervention is a decision that ought to be modeled with an expected utility calculation. I will formulate this insight. Let x be the intended beneficial effect of a drug—say, avoiding a heart attack—which brings utility U(x) to a patient, and let the harmful effects of the drug be y_i, which brings utility $U(y_i)$ to a patient (these are negative). Decision theory holds that, when faced with a decision to take some action or not, in standard cases one should take that action if it would bring more utility than not taking that action, and if it would not, then do not. Of course, any of x and y_i could have occurred without using the drug. Thus the expected utility (EU) of using the drug (D), compared with not using the drug (~D), is:

$$EU_D = [P(x|D) - P(x|{\sim}D)]U(x) + \forall i\sum[P(y_i|D) - P(y_i|{\sim}D)]U(y_i)$$

Note that, among the various outcome measures typically employed in clinical research and described above, the leftmost multiplicand in the leftmost term of EU_D is best-estimated by RD. Indeed, a naïve estimation of P(x|D) would just be based on the frequencies a/(a+b), and a naïve estimation of P(x|~D) would just be based on the frequencies c/(c+d), and since the difference between the former and the latter is just RD, a naïve estimation of the leftmost multiplicand in the leftmost term of EU_D would simply be based on RD. I call this approach naïve for reasons that will become clear in §4 below.[11] Nevertheless, RD is a close estimator of this term required for the expected

[10] If P(E) and P(C) are very similar, say because the study was randomized, then the value of the rightmost multiplicand is entirely determined by the difference of likelihoods, and in (Stegenga, 2013) I argue that comparing likelihoods is the most compelling way to measure the extent to which purported means change the probability of an end—precisely the difference-making capacity that is in question here.

[11] The mere act of consuming a medical intervention can be thought of as a harmful cost, as can the financial cost of the medical intervention. Thus, if one consumes a medical intervention, some of the harmful effects are essentially guaranteed. Other harmful effects—the harmful physiological effects of the drug—have smaller probabilities. If [P(x|D) - P(x|~D)] is only 0.01, say, and for some i of y_i, $P(y_i) = 1$, and some of the harmful effects have a very large disutility U(i), it is not obvious that EU_D is positive. I do not

utility calculation, and compared to relative outcome measures, RD is far superior (relative measures are simply not an option for this estimation).

To illustrate the problem that arises when not taking P(Y) into account with relative measures of effectiveness, consider the drug alendronate sodium (Fosamax), claimed to allegedly cause an increase in bone density in women, used with the aim of decreasing the frequency of bone fractures. A large trial compared the drug to placebo over a four year period (Black et al., 1996). The evidence from the trial was touted as showing that the drug reduces the risk of hip fractures by 50%—this was a relative measure of risk reduction (RRR). However, as Moynihan and Cassels (2005) note, only 2% of the women in the control group had hip fractures during the four years of the trial, while only 1% of the women in the experimental group had hip fractures. Thus the RD effect size was a mere 1%—the absolute difference in hip fracture rates between the experimental group and the control group was only 1%—after consuming the drug for four years. Moreover, it was only women at 'high risk' of hip fractures—namely, those who had already had hip fractures—who were included as subjects in the study, and thus the subjects in the study were not representative of the broader target population of patients for whom such an intervention is intended (which raises the problem of extrapolation, to which I turn in the following section).

If you are confused by the difference between the various outcome measures, then you might maintain the perplexed question: is alendronate sodium effective, or isn't it? After all, we have two outcome measures reporting two effect sizes:

RRR = 50%

RD = 1%

So, does alendronate sodium decrease the chance of hip fractures in the relevant population by 50% or 1%? The answer is that it does both, because the question is ambiguous. For a particular patient, the probability of having a hip fracture after taking alendronate sodium decreases from 2% to 1%, and so, since $2 - 1 = 1$, the chance of having a hip fracture decreases by 1%. But since 1 (of anything) is 50% of 2 (of anything), the probability of having a hip fracture after taking alendronate sodium decreases by 50%. Which effect size should a particular patient and her physician be impressed by? Perhaps both. At the very least, they need the absolute measure to make an informed treatment decision. Effectiveness of an intervention, from the first-person perspective of a patient, is, roughly, the degree to which the intervention increases the probability that the patient will experience the beneficial outcome in question. This difference-making notion is adequately represented by RD and is not adequately represented by RR. So, from an

---

pretend that these numbers are straightforward to calculate. My point is that one ought not assume that the expected utility of using a medical intervention is positive, given only a large relative effect size.

individual patient's perspective, the appropriate outcome measure in this example is RD: the probability of having a hip fracture after taking alendronate sodium decreases by 1%. In short, alendronate sodium is barely effective, even in the most at-risk patients. The use of a relative outcome measure makes the drug seem more effective than it in fact is.

Here is another example of a misleading reliance on relative outcome measures. The Helsinki Heart Study tested the capacity of gemfibrozil to decrease cholesterol levels and thereby decrease cardiac disease and death. After five years of taking the drug, the subjects in the experimental group had a reduced relative risk of cardiac disease of 34%, but since the baseline rate of cardiac disease is so low, this amounted to an absolute reduced risk (RD) of only 1.4% (Frick et al., 1987). Moreover, there was no difference between the groups in the death rate.

The reliance on relative outcome measures at the expense of absolute outcome measures is ubiquitous.[12] This, together with the fact that people overestimate the effectiveness of medical interventions when provided with relative outcome measures, entails that on average people overestimate the effectiveness of medical interventions. Effectiveness always should be measured and reported in absolute terms (using measures such as RD), and only sometimes should effectiveness also be measured and reported in relative terms. This would have the result that estimates of the effectiveness of medical interventions would be more accurately deemed lower than they now are.

One might object that a medical intervention with a low absolute effect size could nevertheless be considered 'effective', because if the medical intervention were used by a large number of people, then a significant absolute number of those people would experience the beneficial outcome of the intervention. This is especially the case with those medical interventions that are widely used today as preventive medications, such as cholesterol-lowering drugs and blood pressure-lowering drugs. For example, if a cholesterol-lowering drug has a 1% absolute reduction in the risk of death, and ten million people consume the drug, then 100 000 lives are saved. That is a great outcome. However, it is not obviously great from the perspective of a particular typical patient. One rationale for the use of such interventions could be similar to the rationale for the use of vaccines: most people who are vaccinated against a certain disease would not have developed the disease in question had they not been vaccinated, and thus they do not directly receive a benefit from the intervention, but the widespread use of vaccines is nonetheless warranted because the practice decreases the overall number of people who develop the disease (thanks to so-called 'herd immunity'). This way of conceiving of the benefits of vaccines requires thinking of the benefit accrued to a population rather than any particular individual. However, that is not how drugs with low effect sizes—those

[12] I am not aware of a careful empirical survey that demonstrates this, but many commentators have noted that risk relativism is widespread. See, for discussion, (Moynihan & Cassels, 2005).

preventive drugs that lower cholesterol, say—should be thought of. An individual patient and her physician want to know that if they employ a particular medical intervention then there is a reasonably good chance that the intervention will be effective for this particular patient. For drugs with low absolute effect sizes like the ones I have been discussing above, that is almost never the case.

An objection related to the one above holds that interventions with high relative effect sizes but low absolute effect sizes are indeed effective—alendronate sodium cuts one's risk of hip fractures in half, after all—it is just that there are relatively few people for whom the intervention can be effective, because the baseline probability of a woman having a hip fracture is so low. If a woman were among the 2% who were going to have a hip fracture, then alendronate sodium would cut that woman's risk in half, which (this objection goes) is significant. There is no reason that a measure of effectiveness should be sensitive to the prevalence of the outcome in question, goes this response, and thus a relative measure is more appropriate than an absolute measure. The trouble with such a response is that one cannot tell in advance if one is in the class of people for whom an intervention might be effective—namely, the class of people who will experience the negative outcome in question. From a particular patient's perspective—one who does not know in advance if she will have a hip fracture, say—a drug like alendronate sodium decreases her chance of having a hip fracture by a tiny amount. Another way of putting the point is: for a particular patient, an intervention with a low absolute effect size is very unlikely to provide any benefit at all.

To see this, consider the absolute outcome measure 'number needed to treat' (NTT). This is an intuitive outcome measure: it tells you how many people would have to use the intervention in question in order to achieve one of the outcomes of interest. If an intervention has an RD of 1%, then the NTT is 100. That is, one hundred people would need to use the intervention in order to achieve one positive outcome. In other words, only 1 of the 100 people who used the intervention would experience the beneficial outcome, while the other 99 would not. There may be other beneficial outcomes of the intervention—perhaps changes in a continuous parameter rather than a dichotomous parameter (but then again, there may be many harms of the drug as well)—but in any case the principle outcome of interest would not be experienced by the vast majority of the people that consume the drug. As above, when deciding whether or not to use a medical intervention, a patient or physician wants to know to what extent would using this particular intervention change the probability of the outcome in question. To determine this, measures like RD or NTT are needed.

As the above examples illustrate, when the base rate of a negative outcome is low, then an intervention employed to avoid that outcome could have a seemingly large relative effect size but a low absolute effect size. Schwartz and Meslin (2008) suggest that the use of absolute measures could cause patients to make irrational decisions (say, to

forgo treatment in cases similar to those above, in which the absolute effect sizes are tiny), and for at least some cases they seem to suggest that this is an argument in favor of the use of relative measures. Their argument is: for a patient to make an autonomous medical decision they must be informed about the extent to which a particular medical intervention is effective; since people display a low degree of numeracy, absolute outcome measures might hinder patients' understanding of effectiveness; thus, employ relative measures. I hope to have shown that such a comparison between people's comparative understanding of relative versus absolute outcome measures is dubious. Relative measures, by promoting the base rate fallacy, fundamentally mislead patients into overestimating effectiveness

The considerations here are concerned with the kinds of outcome measures that should be employed when summarizing data from clinical trials. The point of performing such experiments is to learn something about whether or not (and if so to what extent) a medical intervention will be effective for a broader target population and for a particular patient. Once a trial has been performed and the data from the trial has been analyzed with an appropriate outcome measure, thereby determining an effect size for the intervention, the effect size is used to make an inference about the effectiveness of the medical intervention in a target setting.

4 Extrapolation

A widely held assumption is that the results of clinical trials can be used to directly infer a general capacity of the medical intervention in question. Since the assumption is that the inferred capacity is general, one can infer that the medical intervention would manifest this capacity in a broader population and indeed in any particular patient. For instance, according to some of the leading medical scientists in the evidence-based medicine movement, in order to determine if one can extrapolate the results from clinical trials to a particular patient, one should "ask whether there is some compelling reason why the results should not be applied to the patient. A compelling reason usually won't be found, and most often you can generalize the results to your patient with confidence" (Guyatt & Rennie, 2001) [p. 71]. This is slightly more refined than simple extrapolation—the application of results from a clinical trial to a broader population, and specifically to a particular patient—since the guidance holds that one should determine if there are reasons why one should not extrapolate. Nevertheless, in the same breath the guidance claims that such reasons are rare, and thus the guidance amounts to simple extrapolation,

most of the time, unless one is aware of a countervailing reason. I will call this methodological guidance 'simple extrapolation, unless' (SEU).[13]

Here is another expression of SEU, again from leaders in the evidence-based medicine community: "results of randomized trials apply to wide populations unless there is a compelling reason to believe the results would differ substantially as a function of particular characteristics of those patients" (Post, de Beer, & Guyatt, 2012). Similarly, an epidemiology textbook notes that "generalizing results obtained in one or more studies to different target or reference populations [is] the premier approach that public health professionals and policy makers use" (Szklo & Nieto, 2007). One of the highest profile guidance statements from methodologists in evidence-based medicine (the CONSORT group) re-iterates this view: "therapies (especially drugs) found to be beneficial in a narrow range of patients generally have broader application in actual practice" (Moher et al., 2010). The trouble with this claim is that, ironically, the 'evidence base' for it is extremely thin. Many of the articles that the CONSORT group cites in support of this claim are merely opinion pieces in medical journals; the more rigorous empirical studies that they cite conclude that SEU is in fact problematic. One such article argues that trial design principles "limit the ability to generalize study findings to the patient population" (Gurwitz, Col, & Avorn, 1992), and another claims that "researchers, funding agencies, ethics committees, the pharmaceutical industry, medical journals, and governmental regulators alike all neglect external validity" (Rothwell, 2005). The CONSORT defense of SEU is remarkable for its violation of its own evidence base.[14]

There are a number of problems with SEU. First, it assumes that the default position regarding extrapolation should be that extrapolation is warranted, based on the further assumption that relevant differences between trial subjects and target patients are rare. The 'unless' clause in SEU states a condition, which, if satisfied, overrides the warrant for extrapolation. Post et al. (2012) note several ways in which such an overriding condition could be satisfied, including: if there are pathophysiologic differences in the illness under investigation which could lead to variability in treatment response, if there are differences in a particular patient compared with the experimental subjects that could diminish the treatment response, and if there are differences in patient or physician compliance that could diminish the treatment response. In the passage cited above, Guyatt and Rennie claim that the overriding condition is rarely satisfied—they claim that a particular patient

[13] (Steel, 2007) calls SEU 'simple induction', which he articulates as follows: "Assume that the causal generalization true of the base population also holds approximately in related populations, unless there is some specific reason to think otherwise." My excuse for multiplying terms is that my concern is specifically about extrapolation, and Steel's term is thus slightly misleading. Another bit of terminology: the term of art often used to describe those studies from which extrapolation is warranted is 'external validity'.

[14] The discussion of extrapolation in (Howick, 2011b) seems to have accepted the CONSORT view at face value.

is usually similar in all important respects to the subjects of a trial from which one wishes to extrapolate. However, one of their exception criteria that overrides warrant for extrapolation is the presence of differences between the target patient and the experimental subjects that may diminish the treatment response in the patient. In principle almost any difference between an experimental group of subjects and a target patient may diminish the treatment response.[15] In practice there are always such differences.

Given the large number of criteria that many clinical trials employ which stipulate the properties that a potential subject must have (and other criteria which they cannot have) to be included in the trial, there are almost always differences between a particular real-world patient and the subjects in a clinical trial. Subjects in a clinical trial are virtually never drawn from a random sample of the broader population who have the disease in question, and the criteria that determine eligibility for a clinical trial often render subjects in a trial different in important respects from the broader population of people who have the disease.[16] Some of these differences are liable to modulate the effect of the intervention in question. At the very least, the default assumption should not be that there are no such differences between trial subjects and target patients. For example, in the RECORD trial, which tested the safety of rosiglitazone, there were numerous inclusion and exclusion criteria applied to determine subject eligibility in the trial. The result was that 99% of the subjects in the trial were Caucasian (despite the fact that the trial was performed in dozens of countries), and the subjects in the trial were on average much healthier than the target population. The subjects in the trial had a heart attack frequency of 4.5 per 1000 people per year, which is about 40% of the relevant group (middle-aged people with type-2 diabetes) in the broader population.

The second major problem with SEU is that it is unreliable due to forms of bias that transcend concerns about internal and external validity, such as publication bias. Even if there were in fact no substantial relevant differences between the experimental subjects and target patients—and thus the overriding clause of SEU were not satisfied, and so extrapolation would be warranted according to SEU—the results of published trials from which one was extrapolating could be entirely misleading, because the published trials may represent only a fraction of the trials that were performed. The reason that publication bias is a problem for SEU is that the subset of studies that are published will report a degree of effectiveness which is higher than the degree of effectiveness measured in all relevant studies (including those that are not published).[17] The subjects that are

[15] One of many empirical demonstrations of this is given by Bartlett et al. (2005).

[16] For an insightful formulation of this argument against (Post et al., 2012) see (Fuller, 2013a). One might think that meta-analysis could resolve some of the problems articulated here, but in fact meta-analysis inherits these problems; see (Stegenga, 2011).

[17] This is a widely reported phenomenon. For an example, see (Eyding et al., 2010).

included in published studies differ from the set of all subjects in their degree of responsiveness to the intervention in question, and under the safe assumption that the overall set of subjects (including subjects from unpublished studies) is more representative of target patients than the subset of subjects that are included in published studies with respect to their degree of responsiveness to the intervention, publication bias threatens extrapolation. The presence of publication bias is a threat to any method of extrapolation which does not take the pernicious effects of publication bias into account. But a method of extrapolation could take publication bias into account by decreasing estimates of effectiveness as measured in published studies when predicting the effectiveness of the medical intervention in a target population, and thereby improve on SEU.

A problem with SEU that is closely related to the problem of publication bias is the fact that many results from clinical trials are later overturned by contradictory results. Broadbent (2013) calls this the problem of stability: in epidemiology, notes Broadbent, many research findings are not stable. This is also true in clinical research: many findings that purport to show that a medical intervention is helpful, or purport to show that a medical intervention is not harmful, are contradicted by results from subsequent research. SEU ignores the chance that the present evidence from which one extrapolates will be contradicted by later evidence.

The third major problem with SEU is that it ignores information regarding how the intervention works. Suppose a clinical trial reported that a particular medical intervention has an effect size of x for some specific parameter, but that background knowledge of the mechanisms of action for the medical intervention suggests that it would have a completely different effect incompatible with x. Further suppose that for a particular patient the overriding clause of SEU is not met—there are no reasons to suppose that the patient in question is different in any relevant respects from the experimental subjects of the trial. SEU tells us rather simply to infer that the medical intervention will cause x in this patient. This approach, obviously, disregards the background knowledge of the mechanism of action of this medical intervention.[18]

The reliance on SEU contributes to overestimating the effectiveness of medical interventions, for reasons corresponding to the above problems with SEU. First, the features that real-world patients tend to have that render them different from subjects in clinical trials—compared to subjects in trials, patients tend to be sicker, older, on more medications, and less compliant—usually mitigate treatment response in patients. Second,

[18] Russo and Williamson (2007), and (Steel, 2007), among others, argue that knowledge of the mechanism of action of an alleged medical intervention is useful in warranting causal claims regarding the intervention. This view has been criticized by Howick (2011a) and Broadbent (2013) on the grounds that mechanistic knowledge is not necessary for extrapolation. See also (Illari, 2011) for a clarifying exposition of the Russo-Williamson thesis. All contributors to this debate, nonetheless, appear to agree that knowledge of mechanisms can, at the very least, sometimes aid in extrapolation (though Howick is skeptical of this).

publication bias is asymmetric with respect to estimating effectiveness of medical interventions: trials that suggest a medical intervention is effective are far more likely to be published than trials that suggest a medical intervention is ineffective. Third, attention to the mechanism by which an intervention works ought, at least sometimes, to decrease one's estimate of the effectiveness of medical interventions. In §2 I noted the example of high-dose chemotherapy, which appeared to be effective for breast cancer treatment in small initial trials (and which was already known to be effective for treating blood cancers), but since chemotherapeutic drugs operate by interfering with mechanisms of cell division, cells that divide rapidly (like blood cells) are more susceptible to chemotherapy, and slower growing tumors (like breast tumors) are less susceptible to chemotherapy. The estimation of effectiveness of high-dose chemotherapy for breast cancer, based on seemingly positive results from the initial trials, should have been tempered by consideration of the intervention's mechanism of action.

As if SEU were not problematic enough, the 'unless' clause in SEU is often not attended to in policy development. Fuller (2013b) examined six clinical guidelines that recommend treatment with the five most commonly prescribed classes of medications for elderly patients in Ontario and found that these guidelines employ simple extrapolation for generalizing from RCT results to treatment guidelines for wide target populations (including the elderly), without considering limits to generalizability.

The problems with SEU discussed here naturally suggest three corresponding correctives. First, extrapolation would be more reliable if subjects in trials about a particular disease were more similar to patients in the broader population who have that disease than they currently are. Not just any similarity will do, of course—it would not help if both the experimental subjects and all members of the target population were born under the sign of Scorpio (likewise, as Broadbent (2013) notes, not just any difference between experimental population and target population invalidates an extrapolation). Second, when extrapolating from a particular measured value in published studies to make an inference of effectiveness in a target population, one should incorporate a subtraction factor to account for publication bias (the size of which should correspond to best estimates of the severity of publication bias in the relevant domain). Third, and related to the first, when extrapolating, one ought to consider knowledge of the mechanism of how the intervention works, and how the intervention can fail to work, and one ought to ensure that the target population is similar to the experimental population in all respects that are relevant to these mechanisms.

This latter principle has been characterized in various ways. To know that an intervention that appeared effective in an experimental setting will be effective in a target setting, Cartwright (2011) argues that we must know (i) that the causal law which operated in the experimental setting also operates in the target setting, (ii) that the 'helping factors' (additional causal requirements) which are necessary for the intervention

to be effective (and that were present in the experimental setting) are in place in the target setting, and (iii) that the mechanism by which the intervention is effective operates in the target population and remains unbroken upon application of the intervention. Making a relatively similar point, Broadbent (2013) argues that we must know that we have eliminated potential interferers (where an interferer is a possible way in which an extrapolation could go wrong). Perhaps the most detailed proposal for a compelling method of extrapolation in biology and the social sciences is due to (Steel, 2007), who argues that extrapolation can be grounded in 'comparative process tracing': identifying the relevant mechanism in the experimental population and comparing the mechanism in the target population, and assessing those stages in the mechanism that are most likely to be different between the two populations (although Steel's focus is on extrapolation from experiments on one species to knowledge about another species—typically humans—the method of comparative process tracing can also be valuable in the context of clinical research).

Perhaps most importantly, when extrapolating measurements of effectiveness from a research setting to a clinical setting, one ought to take into account features of the research setting that go beyond mere concern about internal and external validity, including features such as the potential for publication bias and the chance that present findings will be contradicted by later research.

5 Conclusion

The measurement of effectiveness of medical interventions faces three epistemological challenges: the selection of a good measuring instrument, the use of an appropriate outcome measure, and the employment of a reliable method of extrapolating measurements in an experimental setting to a broader setting.[19] The way these challenges are met in contemporary clinical research is unsatisfactory, which systematically contributes to overestimating the effectiveness of medical interventions.

Alexandrova, A. (2008). First person reports and the measurement of happiness. *Philosophical Psychology, 21*(5), 571-583.

Bartlett, C., Doyal, L., Ebrahim, S., Davey, P., Bachmann, M., Egger, M., & Dieppe, P. (2005). The causes and effects of socio-demographic exclusions from clinical trials. *Health Technol Assess, 9*(38), iii-iv, ix-x, 1-152.

[19] There is a second-order issue of measurement in clinical research, namely, the measurement of the quality of evidence for each of these three first-order measurement issues; I address this in (Stegenga, 2014).

Black, D. M., Cummings, S. R., Karpf, D. B., Cauley, J. A., Thompson, D. E., Nevitt, M. C., . . . Ensrud, K. E. (1996). Randomised trial of effect of alendronate on risk of fracture in women with existing vertebral fractures. Fracture Intervention Trial Research Group. *Lancet, 348*(9041), 1535-1541.

Bobbio, M., Demichelis, B., & Giustetto, G. (1994). Completeness of reporting trial results: effect on physicians' willingness to prescribe. *Lancet, 343*(8907), 1209-1211.

Broadbent, A. (2013). *Philosophy of epidemiology*: Palgrave Macmillan.

Brownlee, S. (2008). *Overtreated: why too much medicine is making us sicker and poorer*: Bloomsbury.

Cartwright, N. (2011). Evidence, External Validity, and Explanatory Relevance. *<I>Philosophy of Science Matters: The Philosophy of Peter Achinstein<D>, Morgan, Gregory J (ed)*, 15-28.

Chang, H. (2004). *Inventing Temperature*: Oxford University Press.

Eyding, D., Lelgemann, M., Grouven, U., Härter, M., Kromp, M., Kaiser, T., . . . Wieseler, B. (2010). Reboxetine for acute treatment of major depression: systematic review and meta-analysis of published and unpublished placebo and selective serotonin reuptake inhibitor controlled trials. *Bmj, 341*. doi: 10.1136/bmj.c4737

Forrow, L., Taylor, W. C., & Arnold, R. M. (1992). Absolutely relative: how research results are summarized can affect treatment decisions. *Am J Med, 92*(2), 121-124.

Frick, M. H., Elo, O., Haapa, K., Heinonen, O. P., Heinsalmi, P., Helo, P., . . . et al. (1987). Helsinki Heart Study: primary-prevention trial with gemfibrozil in middle-aged men with dyslipidemia. Safety of treatment, changes in risk factors, and incidence of coronary heart disease. *N Engl J Med, 317*(20), 1237-1245. doi: 10.1056/nejm198711123172001

Fuller, J. (2013a). Rationality and the generalization of randomized controlled trial evidence. *J Eval Clin Pract, 19*, 644-647.

Fuller, J. (2013b). Rhetoric and argumentation: how clinical practice guidelines think. *J Eval Clin Pract, 19*, 433-441.

Gurwitz, J. H., Col, N. F., & Avorn, J. (1992). The exclusion of the elderly and women from clinical trials in acute myocardial infarction. *JAMA, 268*(11), 1417-1422.

Guyatt, G., & Rennie, D. (2001). *User's guide to the medical literature*. Chicago: AMA Press.

Hamilton, M. (1960). A rating scale for depression. *J Neurol Neurosurg Psychiat, 23*, 56-62.

Howick, J. (2011a). Exposing the vanities - and a qualified defense - of mechanistic reasoning in health care decision making. *Philosophy of science, 78*(5), 926-940.

Howick, J. (2011b). *The philosophy of evidence-based medicine*: Wiley.

Illari, P. M. (2011). Mechanistic Evidence: Disambiguating the Russo–Williamson Thesis. *International Studies in the Philosophy of Science, 25*(2), 139-157. doi: 10.1080/02698595.2011.574856

Jensen, P. S., Arnold, L. E., Swanson, J. M., Vitiello, B., Abikoff, H. B., Greenhill, L. L., . . . Hur, K. (2007). 3-year follow-up of the NIMH MTA study. *J Am Acad Child Adolesc Psychiatry, 46*(8), 989-1002. doi: 10.1097/CHI.0b013e3180686d48

Lavery, A. M., Verhey, L. H., & Waldman, A. T. (2014). Outcome Measures in Relapsing-Remitting Multiple Sclerosis: Capturing Disability and Disease Progression in Clinical Trials. *Multiple Sclerosis International, 2014*, 13. doi: 10.1155/2014/262350

McClimans, L. (2010). A theoretical framework for patient-reported outcome measures. *Theoretical Medicine and Bioethics, 31*, 225-240.

McClimans, L. (2013). The role of measurement in establishing evidence. *Journal of Medicine and Philosophy, 38*(5), 520-538.

Moher, D., Hopewell, S., Schulz, K. F., Montori, V., Gøtzsche, P. C., Devereaux, P. J., . . . Altman, D. G. (2010). *CONSORT 2010 Explanation and Elaboration: updated guidelines for reporting parallel group randomised trials* (Vol. 340).

Molina, B. S., Hinshaw, S. P., Swanson, J. M., Arnold, L. E., Vitiello, B., Jensen, P. S., . . . Houck, P. R. (2009). The MTA at 8 years: prospective follow-up of children treated for combined-type ADHD in a multisite study. *J Am Acad Child Adolesc Psychiatry, 48*(5), 484-500. doi: 10.1097/CHI.0b013e31819c23d0

Moynihan, R., & Cassels, A. (2005). *Selling sickness : how the world's biggest pharmeutical companies are turning us all into patients*. Vancouver: Greystone Books.

Naylor, C. D., Chen, E., & Strauss, B. (1992). Measured enthusiasm: does the method of reporting trial results alter perceptions of therapeutic effectiveness? *Ann Intern Med, 117*(11), 916-921.

Nexøe, J., Gyrd-Hansen, D., Kragstrup, J., Kristiansen, I. S., & Nielsen, J. B. (2002). Danish GPs' perception of disease risk and benefit of prevention. *Fam Pract, 19*(1), 3-6. doi: 10.1093/fampra/19.1.3

Post, P. N., de Beer, H., & Guyatt, G. H. (2012). How to generalize efficacy results of randomized trials: recommendations based on a systematic review of possible approaches. *J Eval Clin Pract*. doi: 10.1111/j.1365-2753.2012.01888.x

Rothwell, P. M. (2005). External validity of randomised controlled trials: "to whom do the results of this trial apply?". *Lancet, 365*(9453), 82-93. doi: 10.1016/s0140-6736(04)17670-8

Russo, F., & Williamson, J. (2007). Interpreting causality in the health sciences. *International Studies in the Philosophy of Science, 21*, 157-170.

Schachter, H. M., Pham, B., King, J., Langford, S., & Moher, D. (2001). How efficacious and safe is short-acting methylphenidate for the treatment of attention-deficit disorder in children and adolescents? A meta-analysis. *Can Med Assoc J, 165*(11), 1475-1488.

Schwartz, P. H., & Meslin, E. M. (2008). The ethics of information: absolute risk reduction and patient understanding of screening. *J Gen Intern Med, 23*(6), 867-870. doi: 10.1007/s11606-008-0616-y

Sorensen, L., Gyrd-Hansen, D., Kristiansen, I. S., Nexoe, J., & Nielsen, J. B. (2008). Laypersons' understanding of relative risk reductions: randomised cross-sectional study. *BMC Med Inform Decis Mak, 8*, 31. doi: 10.1186/1472-6947-8-31

Steel, D. (2007). *Across the boundaries: extrapolation in biology and the social sciences*. New York: Oxford University Press.

Stegenga, J. (2011). Is Meta-Analysis the Platinum Standard? *Stud Hist Philos Biol Biomed Sci, 42*, 497-507.

Stegenga, J. (2013). Probabilizing the end. *Philosophical Studies, 165*, 95-112. doi: 10.1007/s11098-012-9916-5

Stegenga, J. (2014). Quality of Information in Clinical Research. In P. M. Illari & L. Floridi (Eds.), *The Philosophy of Information Quality*: Springer.

Stegenga, J. (forthcoming). Effectiveness of Medical Interventions. *Studies in the History and Philosophy of Biological and Biomedical Sciences*.

Szklo, M., & Nieto, J. (2007). *Epidemiology: beyond the basics* (2nd ed.). Boston: Jones and Bartlett Publishers.

Tal, E. (2011). How accurate is the standard second? . *Philosophy of science, 78*(5), 1082-1096.

Tal, E. (forthcoming). Making Time: A Study in the Epistemology of Measurement. *British Journal for the Philosophy of Science*.

Teller, P. (2013). The concept of measurement-precision. *Synthese, 190*, 189-202.

van Fraassen, B. (2008). *Scientific representation: paradoxes of perspective*. New York: Oxford University Press.

Worrall, J. (2010). Do we need some large, simple randomized trials in medicine? In M. Suarez, M. Dorato, & M. Redei (Eds.), *EPSA Philosophical Issues in the Sciences*: Springer Netherlands.