Studies in History and Philosophy of Biological and Biomedical Sciences 54 (2015) 62–71

Contents lists available at ScienceDirect

Studies in History and Philosophy of Biological and Biomedical Sciences

journal homepage: www.elsevier.com/locate/shpsc

Measuring effectiveness

Jacob Stegenga [a,b,*]

[a] Department of Philosophy, University of Victoria, Canada
[b] University of Johannesburg, South Africa

ARTICLE INFO

Article history:
Available online 18 July 2015

Keywords:
Measurement
Effectiveness
Evidence
Outcome measures
Extrapolation
RCT

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.

When citing this paper, please use the full journal title *Studies in History and Philosophy of Biological and Biomedical Sciences*

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

[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.

* Department of Philosophy, University of Victoria, Canada.
E-mail address: stegenga@uvic.ca.

http://dx.doi.org/10.1016/j.shpsc.2015.06.003
1369-8486/

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 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 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.

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