Studies in History and Philosophy of Biological and Biomedical Sciences 54 (2015) 49–61

Contents lists available at ScienceDirect

Studies in History and Philosophy of Biological and Biomedical Sciences

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

The Risk GP Model: The standard model of prediction in medicine

Jonathan Fuller [a,*], Luis J. Flores [b]

[a] Faculty of Medicine, University of Toronto, Canada
[b] Department of Philosophy, King's College London, United Kingdom

ARTICLE INFO

Article history:
Available online 26 July 2015

Keywords:
Prediction
Epidemiology
Medicine
Risk
Extrapolation
Probability

ABSTRACT

With the ascent of modern epidemiology in the Twentieth Century came a new standard model of prediction in public health and clinical medicine. In this article, we describe the structure of the model. The standard model uses epidemiological measures—most commonly, risk measures—to predict outcomes (prognosis) and effect sizes (treatment) in a patient population that can then be transformed into probabilities for individual patients. In the first step, a risk measure in a study population is *generalized* or extrapolated to a target population. In the second step, the risk measure is *particularized* or transformed to yield probabilistic information relevant to a patient from the target population. Hence, we call the approach the Risk Generalization–Particularization (Risk GP) Model. There are serious problems at both stages, especially with the extent to which the required assumptions will hold and the extent to which we have evidence for the assumptions. Given that there are other models of prediction that use different assumptions, we should not inflexibly commit ourselves to one standard model. Instead, model pluralism should be standard in medical prediction.

© 2015 Elsevier Ltd. All rights reserved.

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

1. Introduction

Predictions are central to medical practice. Doctors want to know what will happen to the patient in the future given their present condition (prognosis), and how treatment or prevention might alter the natural course of events (intervention). But is there a standard model of prediction in medicine, a dominant approach in which trainees are schooled and according to which doctors practice? What we are after is a prediction scheme similar to other models of prediction in the philosophy of science, the most classic and well-known of which is the Deductive-Nomological Model (DN Model) of Carl Hempel and Paul Oppenheim (Hempel & Oppenheim, 1948).[1]

Such an idealization is not to be found in medical textbooks. In fact, textbooks tend not to use the term 'prediction' to label a major category of clinical inference, but instead divide inferential activities into the traditional medical categories of diagnosis, prognosis, therapy and harm (Guyatt, Rennie, Meade, & Cook, 2008). Yet prognostic, therapeutic and harm-related inferences typically involve predictions, hypotheses about future outcomes. Even diagnosis can be conceptualized as a predictive activity; clinical textbooks speak of "clinical prediction rules" for diagnosis and the "positive predictive value" of a diagnostic test (Guyatt et al., 2008, pp. 491–505; Fuller, Sankar, & Upshur, 2013, pp. 580). Diagnosis is predictive in the wider sense of inferring an outcome that is not definitively known (i.e. the presence of a particular disease). It will be profitable to examine the shared structure of these distinct types of clinical inference.

An important clue to the existence of a standard model is that there seems to be a common target of several critiques of medical prediction, some of which will be explored in Sections 4 through 6. However, the received model lacks an explicit philosophical reconstruction—or a reconstruction of any sort, for that matter. Without a clear representation, it remains a nebulous target.

Abbreviations: GP, Generalization–Particularization; EBM, evidence-based medicine; AR, absolute risk; CVD, cardiovascular disease; ES, effect size; RR, relative risk; RD, risk difference; NHS, National Health Service.

* Corresponding author.
E-mail address: jonathan.fuller@mail.utoronto.ca (J. Fuller).

[1] Another notable model of prediction is found more recently in the work of Spirtes, Glymour, and Scheines (2000). Their approach uses directed graphical modelling to predict the probability distribution resulting from a targeted intervention on one or more variables.

http://dx.doi.org/10.1016/j.shpsc.2015.06.006
1369-8486/© 2015 Elsevier Ltd. All rights reserved.

Here, we reconstruct and examine the Risk Generalization–Particularization (Risk GP) Model, the standard model of prediction in medicine. Risk GP is standard in that it represents the dominant prescriptive model in contemporary practice (the gold standard), as well as the model that many practitioners implicitly rely upon when making evidence-based decisions. Risk GP is an epidemiological model, relying centrally on aggregate outcomes in populations. Like the science of epidemiology, the model is relatively new when framed against the long history of medicine, although rational approaches to prediction have been around since at least the time of Hippocrates (460-370 BCE). The Risk GP Model actually consists of two inferences in series: a *generalization* of a risk measure from a study population to a target patient population of interest; and a *particularization*, a transformation of this measure to yield probabilistic information about a patient within the target population.

There are well-known problems at both stages. Most worryingly, the necessary assumptions for generalization and particularization may not hold widely, and even when they do hold, we might not have evidence to warrant them. These problems are not an inevitable challenge for clinical practice, or even for epidemiological predictions, but are peculiar to the Risk GP Model. Of course, most models are imperfect, and their ideal assumptions will sometimes fail to represent reality. Those circumstances demand flexibility; we should not commit ourselves to a one-model-fits-all approach, but should be model pluralists instead.

2. Models of prediction in historical perspective

A few distinctions will be useful upfront. Alex Broadbent identifies a "process/product" ambiguity in the concepts of prediction. He distinguishes two senses of the term: prediction as a *claim*, and prediction as an *activity* (2013, pp. 86, 89–93). The first sense of 'prediction' is a claim or hypothesis, such as: 'it will rain tomorrow'. The second sense of 'prediction' is an activity or argument, such as: 'following many previous weeks like this one it rained the next day; therefore, it will rain tomorrow'. Prediction activities are inferences involving prediction claims.

In cases like the meteorological prediction just mentioned, prediction activities are inferences with a prediction claim, a definite forecast, as their conclusion. We can call these prediction activities *predictive inferences* to distinguish them from prediction activities that do *not* have a definite prediction claim as their conclusion. For instance, take the inference: 'on 60% of previous weeks like this one it rained the next day; therefore, the probability that it will rain tomorrow is 60%'. The conclusion is not a prediction claim; in asserting it, we are not placing a bet or committing ourselves to the occurrence of some future event. If instead it was clear skies without any approaching storm fronts, the meteorologist would conclude that the probability of rain is low, which is obviously not a prediction that it will rain tomorrow. Yet we might still want to call this statistical inference a 'prediction activity' because the conclusion tells us the probability *of* a prediction claim.[2]

As previously alluded, there are at least two, non-exclusive types of prediction claims in natural language and medical discourse. The more inclusive type encompasses all hypotheses about unknown (unobserved) events or outcomes. It includes diagnostic hypotheses like: 'the patient has heart disease'. Meanwhile, the less inclusive type of prediction claim is a subtype of the former, and includes only hypotheses about the future (e.g. 'the patient will experience a cardiovascular disease event over the next ten years'). In these cases, the outcomes are unknown specifically because they have not yet occurred (Broadbent calls this less inclusive type "narrow prediction" (2013, pp. 93)). As we will see, the standard model can account for predictions in the broader sense. But since prognostic and therapeutic predictions are usually predictions in the narrow sense (they are hypotheses about what will happen in the future to the patient), narrow predictions will be our main focus.[3]

An informative prediction scheme would model both the prediction claims and the associated prediction activities in a given field. The DN Model (Hempel & Oppenheim, 1948) provides a good illustration. Given a physical phenomenon to be explained (the explanandum), we supply the laws of nature and particular facts that jointly entail it (the explanans). To explain why an object is accelerating at a particular rate of 1/2 m/s^2, we can deduce the rate from Newton's Second Law and some initial conditions:

$$\begin{array}{c}\text{Acceleration} = \text{Force/Mass} \\ \underline{\text{Force} = 1\text{ N, Mass} = 2\text{ kg}} \\ \text{Acceleration} = 1/2\text{ m/s}^2\end{array}$$

In the DN scheme, explanation and prediction are symmetrical activities; a prediction is an explanation in which the explanans (above the line) is known but the explanandum (below the line) is not. So the DN Model is also a model of prediction in the wide sense. The entire model represents a prediction activity, while the conclusion represents a prediction claim about an unknown variable.

Unfortunately, the DN Model is of limited use in characterizing modern medical prediction.[4] Few universal laws are used in clinical practice, and there is no unifying theory akin to Newton's Laws. Yet the absence of any grand theory in contemporary medicine is peculiar from a historical perspective. The miasma and contagion theories of disease persisted well into the Nineteenth Century (Gillies, 2005), and from Ancient Greek medicine until the Renaissance, the Hippocratic Theory of the Four Humours, a paradigmatic example of a unifying medical theory, provided a theoretical basis for medicine (Duffin, 2010, pp. 42–45).

In the canonical interpretation of humoural theory, the balance of four bodily fluids—blood, phlegm, black bile and yellow bile—determines a person's state of health or disease. When each of the four humours is in equilibrium, the person is healthy; when any are in disequilibrium, the person is diseased. It follows that reversing disequilibrium in disease restores health. Thus, for bilious patients (with excess bile) and phlegmatic patients (with excess phlegm) the Hippocratic *Affectations* makes the following prescription: "In cleaning, employ medications according to the following principle: when patients are bilious, give medications that clean out bile; when they are phlegmatic, give medications that clean out phlegm" (Potter, 1988, pp. 43). Bloodletting, a therapy commonly used for thousands of years and for a wide range of

[2] Predictive inferences in medicine are typically also 'probabilistic' in that they warrant the definite prediction claim inductively. The essential difference is that a predictive inference concludes that the outcome or event will occur, while this statistical inference merely derives the probability of its occurrence.

[3] Narrow prediction claims include subjunctive conditionals, or 'counterfactuals' ('if T, then O'); specifically, counterfactuals in which the consequent refers to some future outcome or event. In order to decide on a course of action, especially when multiple alternative courses are open, physicians must often predict what will happen before the antecedents of the outcome are established. For instance, what will happen in the future to the patient *if* they are treated?

[4] Hempel (1962) also proposed a statistical model analogous to the DN Model that at first glance might seem more relevant, but because the model is intertwined with Hempel's interpretation of probability we will not discuss the details of the statistical model here.

diseases, was initially founded on a similar principle (Duffin, 2010, pp. 72).

We can model this kind of theoretical or mechanistic prediction as follows.

The quantity/quality of a humour is imbalanced to degree *x* (disease)
Intervention: *-x*
The quantity/quality of the humour is balanced (health)

The humoural scheme is a model of prediction for treatment that follows the DN pattern. The prediction activity represented by the model is a predictive inference because we directly infer the prediction claim.

Aside from humoural theory, Hippocratic medicine was also notable for its emphasis on clinical observation. Hippocrates and his followers realized that diseases have a predictable natural course that can be abstracted from observation of similar cases. The Hippocratic *Prognostic* dictates that: "He who would make accurate forecasts as to those who will recover, and those who will die, and whether the disease will last a greater or less number of days, must understand all the symptoms thoroughly…For it is by the same symptoms in all cases that you will know the diseases that come to a crisis at the times I have stated" (Potter, 1988, pp. 48).

We can represent the reasoning as follows.

Patients *a*, *b*, *c* had fever lasting more than *x* days and then died
Patient *d* has fever lasting more than *x* days
Patient *d* will die

In contrast to the humoural scheme for treatment, the clinical observation scheme is a model of prediction for prognosis. Greek for 'the process of knowing before', prognosis is the act of predicting a future clinical outcome in a treated or untreated case. Once again, the prediction activity is a predictive inference. However, rather than deduction from theory, this time the inference is an enumerative induction from past experience.[5]

Throughout subsequent history, the two distinct kinds of prediction that coexisted in Hippocratic medicine—inference from theory and induction from experience—continued to coexist, though sometimes as rival rather than complementary approaches. Robyn Bluhm and Kirstin Borgerson describe "two traditions in medicine": the "rationalist" and "empiricist" traditions (2011, pp. 204). Inference from theory is most aligned with the rationalist or mechanist tradition in the history of medicine. The mechanist tradition championed the establishment of medical theory or mechanisms of clinical causation as the route to medical knowledge; we can identify monumental figures like Paracelsus (1493–1541) and Claude Bernard (1813–1878) with this strain. In comparison, induction from experience is better linked to the empiricist tradition, which emphasized observation, classification, counting, and comparison, and within which Thomas Sydenham (1624–1689) and P.-C.-A. Louis (1787–1872) can be placed.

Several antecedents were important in the rise of the standard model of contemporary medical prediction, which grew out of the empiricist tradition. With the development of epidemiology and statistics in the early Twentieth Century came new methods for classifying, counting and comparing, as well as for inductive inference. Then in the 1960s, the science of *clinical* epidemiology was born, which sought to apply these new tools to clinical research (Bluhm & Borgerson, 2011). Through carefully quantifying outcomes in a population of patients, one can predict outcomes or determine their probability in future patients in a more precise way than is offered by simple enumerative induction.

While early Twentieth Century remedies typically aimed to cure disease or relieve symptoms, the second half of the century saw a shift towards disease prevention, especially the prevention of cancer and heart disease, the leading killers in developed societies (WHO, 2011). Making use of their new tools, epidemiologists and clinical researchers began to classify risk factors, quantify the risk of disease associated with these factors, and measure the effectiveness and safety of new interventions to prevent disease.

Meanwhile, a new era of consumer protectionism was beginning, including in healthcare. In the USA, the federal government granted increasing powers to the Food and Drug Administration (FDA) to regulate the pharmaceutical marketplace, eventually requiring that all new drugs demonstrate safety and efficacy in clinical trials prior to market approval (Peltzman, 1973). This legislation guarantees the existence of population studies for all new drugs marketed in the USA. However, in the early 1990s a group of clinical epidemiologists primarily located at McMaster University in Canada began to express concern that despite the abundance of tools from clinical epidemiology as well as epidemiological and clinical research evidence, medical practitioners were not using the tools or keeping up with the evidence. From their efforts, the evidence-based medicine (EBM) movement sprung into existence.

EBM represents the maturation of the medical empiricist tradition; it advocates applying the results of population studies over mechanistic reasoning or induction from personal clinical experience in diagnosis, prognosis and therapy (Howick, 2011a). It was defended as a new "paradigm" for medical practice (Djulbegovic, Guyatt, & Ashcroft, 2009; EBM Working Group, 1992), and has become dominant in medical research and education, accepted by leading medical schools and all of the major medical journals. With the increase in evidence-based clinical practice guidelines late in the Twentieth Century in terms of number and influence (Upshur, 2014), EBM set the new standard for clinical reasoning.

These developments led to the emergence and ascendance of a new model of medical prediction, which we will introduce in the next section.

3. Introducing the Risk GP Model

Two approaches—inference from theory and induction from experience—characterized medical prediction from Hippocratic medicine down to the present. In the Twentieth Century, owing to the development of clinical epidemiology, a shift towards disease prevention, tighter drug regulation, the emergence of the EBM movement, and an increase in practice guidelines, a new epidemiological model—the Risk GP Model—emerged. A refinement on the 'induction from experience' approach, it has become the standard model of prediction in contemporary medicine.

Part of the advantage of an epidemiological or statistical model of prediction over simple enumerative induction is that patients with the same observed clinical features develop different outcomes. One patient with high blood pressure will have a heart attack, while another patient will not. This difference might arise because the causal factors responsible for heart attack in a patient with high blood pressure are indeterministic. The difference might also occur because patients with high blood pressure vary in factors (including unknown factors) that contribute to heart attack.

As a result of differing outcomes among patients, the frequency of an outcome in a population will rarely be zero or one. Nonetheless, public health practitioners may wish to predict the frequency in order to plan health services or enact health policy. To

[5] The models we have presented are necessarily idealizations of the Hippocratic approach; 'necessarily' because all models in the present sense are—by definition—ideal representations.

these ends, we can measure the frequency of the outcome in a study population and use the result to predict the future frequency in a particular target population.

Population-level predictions are only indirectly relevant for bedside medicine because physicians treat individual patients, not populations. The physician and their patient want to know the particular patient's prognosis and whether the patient will respond to the proposed treatment. Reasoning about the mechanism that produces the outcome in order to answer these questions undoubtedly faces challenges.[6] We do not know in exactly which patients the mechanism will operate successfully to produce the outcome, or else we would define our target populations more narrowly than at present to include just those individuals. Of course, neither does the frequency of the outcome in the target population tell us which patients will develop the outcome. But this frequency *can* help us to quantify our uncertainty; we can determine the probability that a patient in the target population will develop the outcome.

Frequencies are not the only aggregates we can measure in a population. We could also calculate the average for a clinical variable with a range of values greater than two, and often do. But frequencies are sometimes easier to measure and often easier to interpret. Furthermore, frequencies—but not averages—can tell us the probability of the outcome for a patient. Perhaps for these reasons, frequencies are so often used in medical prediction, and will serve as our paradigm example of an aggregate outcome. The *absolute risk* (AR) is the epidemiological term for the frequency of outcome *O* in a population:

$$\mathrm{AR} = (\text{number of } O \text{ in population})/(n \text{ of population})$$

The absolute risk figures in prognostic prediction claims. For instance, according to the widely used Framingham Risk Scale, the ten-year untreated risk ($\mathrm{AR}_{\neg X}$) of cardiovascular disease (CVD), including a heart attack or stroke, in a woman with 19 'CVD points' (assigned based on a number of risk factors) is about 25% (D'Agostino et al., 2008). In other words, we *predict* that in a population of untreated women with 19 CVD points, 25% will develop a CVD outcome over the next ten years.

In patients that are part of a "high-risk" population, defined as a population in which the untreated risk of CVD is 20% or greater (Genest et al., 2009, pp. 570), a cholesterol-lowering drug called a statin is often prescribed. From clinical trial evidence (Baigent et al., 2005) we might predict that the *treated* risk (AR_X) in the high-risk group described above—the frequency of CVD among patients treated with a statin—is approximately 20%.[7] This figure alone tells us nothing about the effectiveness of the drug; the untreated risk might be the 25% cited above, but it might also be 20% (the treatment prevents no CVD on net) or even <20% (the treatment *causes* CVD on net). To discover a treatment's overall effectiveness, we must compare the AR_X to the $\mathrm{AR}_{\neg X}$ using some measure of effect size (ES) such as the *relative risk* (RR)[8]:

$$\mathrm{RR} = \mathrm{AR}_X/\mathrm{AR}_{\neg X}$$

Measures of effect size are used in therapeutic (effectiveness or harm) prediction claims. For example, the relative risk of a CVD event due to statin therapy is 0.8 (RR = 20%/25%); as an effect of the statin, we predict that four fifths as many treated patients will have a heart attack or stroke over the next ten years compared to untreated patients.

We could also quantify the effect size using the *absolute risk reduction*, also known as the *risk difference* (RD):

$$\mathrm{RD} = \mathrm{AR}_{\neg X} - \mathrm{AR}_X$$

The risk difference of CVD due to statins in women with 19 CVD points is 5% (RD = 25%−20%); in other words, we predict that the difference in frequency of CVD events between treated and untreated patients attributed to the statin will be 5%.

In short, predictions in the standard model involve *risk measures*: either the absolute risk or a measure of effect size derived from the absolute risk such as the relative risk or risk difference. For prognostic predictions (involving the absolute risk), the standard model can be represented as follows.

Prognosis:
In the study population, AR = *r* for *O*
...
In target population *F*, AR = *r* for *O*
...
For patient *a* in target population *F*, p(*O*|*F*) = *r*

From the absolute risk of outcome *O* measured in a study population, we predict the absolute risk in a target population defined by clinical features 'F', which can be filled in with whatever prognostic factors we choose. From this prediction claim, we then infer the probability of *O* for a patient from the target population. The prediction activity represented by the model is thus a serial inference consisting of two sub-inferences. As in all serial inferences, the conclusion of the first sub-inference (the first three lines) is a premise in the second sub-inference (the final three lines). Furthermore, both sub-inferences are enthymematic because they rely upon other premises or assumptions (represented by the ellipses) that are suppressed. As we will see, different assumptions might validate the model equally well; but its core structure—the expressed premises and the ultimate conclusion—is unchanging.

As an illustration, in determining that the patient we described earlier has an untreated CVD risk of 25%, we implicitly relied on this model. Our judgement required that we place the patient in a well-defined target or reference population (women with 19 CVD points) that has an associated absolute risk of 25%, as reported in the clinical literature (Genest et al., 2009). 'Women with 19 CVD points' describes a very large (perhaps temporally unbounded) population. The researchers who developed the Framingham prognostic model certainly did not predict the absolute risk in this population by measuring properties of each of its members. Rather, our 25% absolute risk prediction is an extrapolation from a study that followed a sample of women (and men) over time and measured the frequency of cardiovascular disease events that accrued (D'Agostino et al., 2008).

Within the Risk GP Model, *O* can refer to a future event or outcome (narrow prediction), or it can refer to any unobserved event or outcome (wide prediction). The prediction activity that leads me to call 'heads' when the coin is in the air is just as sound if the coin has already landed but is covered by your hand. Thus, the version of Risk GP presented above can also be used for diagnostic predictions. In diagnosis, the outcome has already developed (several patients in the target population already have the disease *O*) but the frequency is not definitively known. Rather than

[6] We will not examine them in detail here, but see Howick (2011b) and Clarke et al. (2014) for a discussion of some of the challenges.

[7] The figure cited is for a 1 mmol/L reduction in LDL cholesterol, which will be assumed from this point on. Therapeutic effectiveness is expected to vary by the amount of reduction in LDL cholesterol achieved.

[8] 'Effect size' sometimes refers exclusively to measures in which the difference in outcome between groups is divided by the variability (Guyatt et al., 2008, pp. 782). Here, we use the term to denote any measure of causal association between exposure and outcome.

prognostic factors, 'F' now refers to signs, symptoms and diagnostic test results.

For therapeutic predictions, the standard model is essentially the same, but involves the effect size rather than the absolute risk.

Therapy (Effectiveness or Harm):
In the study population, ES = r for O,X

In target population F, ES = r for O,X

For patient a in target population F, $\Delta p(O|F) = r$

We start with the effect size for an outcome O due to exposure X (a treatment), measured in a study population. From the effect size in the study population, we predict the effect size in target population F. Then from the effect size in the target population, we derive the change in probability of O for a patient from the target population upon exposure to X.

For example, we might want to treat our patient who has a 25% untreated risk of CVD with a statin drug. The medical literature (Baigent et al., 2005) reports that we can reduce the patient's risk of CVD by one fifth in relative terms (from 25% to 20%). Recall that the untreated risk of CVD varies by reference population, and is not 25% in all patients. Thus, in order to derive the 25%-to-20% risk change we again had to place our patient in the target population consisting of women with 19 CVD points. The treatment effect size we predict for this target population is extrapolated from clinical trials enrolling a much smaller study population.

In summary, standard prediction in medicine is a two-stage process. The first stage or sub-inference is a *generalization* or extrapolation of a risk measure from a study to a target population of our choice. The second stage or sub-inference is a *particularization* or transformation of the value of the risk measure to yield probabilistic information about a patient from the target population. Hence, we call the standard model the *Risk Generalization–Particularization (Risk GP) Model* (the abbreviation is fitting considering that most medical risk management takes place in general practice (GP)). Generalization can be seen as the epidemiology, public health or clinical practice guideline stage because the conclusion is a population-level prediction. On the other hand, particularization is best seen as the clinical medicine stage because the conclusion concerns particular patients.

To be sure, the Risk GP Model is not the only prediction model used in medicine. Enumerative induction is still important in diagnosis and prognosis as physicians develop clinical experience and an acute sensitivity to patterns of disease progression. It may also underlie the prediction that a patient will benefit from a treatment or experience certain side effects based on their previous response to the treatment. On the other hand, mechanistic reasoning resembling the humoural scheme may be used to justify treatment for diseases of excess or deficiency.

Yet the Risk GP Model is aptly considered the standard model because it is held up as a normative ideal by medical authorities, if sometimes only implicitly. In other words, it is the gold standard. The *Users' Guides to the Medical Literature*, an authoritative evidence-based medicine textbook, claims that applying study results to patient care requires asking whether you can "*generalize*" or "particularize" the results to your patient (Guyatt et al., 2008, pp. 6). Paul Glasziou and David Mant explain what this assessment involves in the context of treatment decisions: "The first stage involves an assessment of the transferability of the trial evidence [to your care setting]; the second deals with the application by the clinician to the individual" (2007, pp. 88–89).

Other textbooks, surveys and articles also identify generalization or particularization as the gold standard approach (Djulbegovic, Hozo, & Greenland, 2011; Fuller, 2013a; Glasziou & Irwig, 1995; Goodman, 1999; Post, de Beer, & Guyatt, 2013; Straus, Glasziou, Richardson, & Haynes, 2011; Szklo & Nieto, 2007, p. 376). For instance, Piet Post et al. (2013) did a systematic review of the medical literature to identify approaches to generalizing efficacy results. They recommended that decision-makers generalize the relative effect size (e.g. relative risk) found in randomized clinical trials to the target patient population, a proposal that was reflected in most of the sources they identified, including several EBM guides. In order to determine how a treatment will lower a particular patient's risk, we then need to consider their particular untreated risk, which we infer from a prognostic study (Glasziou & Irwig, 1995; Glasziou & Mant, 2007).

Moreover, Risk GP is the inferential model implicit in evidence-based practice guidelines, which set the *standard* of medical care. Jonathan Fuller (2013a) surveyed clinical guidelines recommending the use of common prescription medications and found that generalization from clinical trial results was used every time to support treatment recommendations for a guideline's target patient population. Clinicians who follow the advice of evidence-based guidelines thus rely—if unknowingly—on the soundness of an extrapolation inference. They may not always interpret the effect sizes reported in guidelines probabilistically, but in the case of preventive treatment, the benefits of the recommended therapy are often described as a lowering or reduction of a patient's risk of the undesired outcome (Genest et al., 2009; Papaioannou et al., 2010; Rabi et al., 2011); the use of this language promotes a probabilistic interpretation. By establishing the standard of care, clinical guidelines play a substantial role in directing clinical practice; this is especially the case when the standard is reinforced through schemes that reward providers for following guideline recommendations or through computerized prompts that encourage adherence to recommendations at the point of care. Through authoritative guides and practice guidelines, Risk GP is a powerful standardizing influence on modern clinical practice.

The Risk Generalization–Particularization Model is the new standard. It tries to overcome the challenges with other models of prediction by using outcomes in study populations as a basis for prediction activities involving target populations and individual patients. In the next two sections, we will explore each constitutive inference—generalization and particularization—in turn. We will also rehearse some old problems and raise some new concerns with the standard model.

4. Generalization

4.1. The inference scheme

In the study population, AR = r or ES = r

In target population F, AR = r or ES = r

Medicine (broadly construed to include public health) makes population-level predictions. For instance, the World Health Organization predicts that the number of annual worldwide cardiovascular deaths will be 23.3 million in 2030 (Mathers & Loncar, 2006; WHO, 2011). Prognostic predictions are typically based on the frequency of the outcome measured in study populations. Thus, the Framingham Heart Study measured cardiovascular disease outcomes over twelve years in a cohort of 8,491 adults with no history of major CVD (D'Agostino et al., 2008). It was from these

data that the Framingham Risk Scale, an algorithm to predict the absolute risk in targeted risk groups, was developed.[9]

Similarly, in treatment we use the Risk GP Model to predict the effect size in the target from the effect size in a study. Rather than the frequency or average value of the outcome, the effect size quantifies the *change* in the frequency or average. Our earlier effect size prediction for statin therapy was based on the relative risk reported in the Cholesterol Treatment Trialist meta-analysis of randomized controlled trials, which analyzed data from 90,056 trial participants (Baigent et al., 2005). The inference involved in predicting the effect size is variously called a 'generalization', 'extrapolation' or 'transportation' of the treatment effect (Horton, 2000). We can represent the extrapolation of either the absolute risk or the effect size using the inference scheme above. We can also identify the scheme as a predictive inference because we directly infer a prediction claim about a risk measure in a target population. Of course, we should not demand that this prediction is perfectly accurate because we never expect that the risk measure in the target will *exactly* equal the risk measure in the study (perhaps the true relative risk is 0.8 in the study and 0.85 in the target). The prediction succeeds so long as the absolute risk or effect size we predict by extrapolating from the study is *close enough* to the true value in the target (though we usually do not specify how close is 'close enough').

Treatment effect predictions depend upon the prior judgement that the association we are extrapolating is indeed an *effect* of the treatment rather than a correlation with a different explanation. This prior 'causal inference' is distinct from the 'causal prediction' that we have so far been considering. The causal inference begins with a correlation between exposure and outcome of magnitude r in the study population; it concludes that the correlation is the effect of the exposure in the study. In contrast, the causal prediction starts from this effect size of magnitude r in the study, and concludes that the effect size in a target population will also be r.

Interpreting the effect size, telling a causal story based on the numbers, is more difficult than it first appears. Recall that we predict a CVD risk difference of 5% due to a statin medication in a population with an untreated risk of 25% ($RD = AR_{\neg X} - AR_X = 25\% - 20\% = 5\%$). It is tempting to conclude that the statin will prevent CVD in exactly 5% of treated patients, but this conclusion does not follow from the numbers alone. As Broadbent argues (2013, pp. 117–121), an exposure may cause or prevent the effect with a *greater* frequency than is quantified by the measure.

This scenario can occur if the exposure replaces other causes that would have otherwise produced the outcome. In a study examining the effects of one hour/day exercise on quality of life, busier participants might substitute the exercise for another activity they were previously enjoying. The exercise regime might cause wellbeing for these participants but in lieu of their previous activities, so there would be no difference in overall effect to quantify. Furthermore, as is well recognized, a 5% net difference in outcome is perfectly consistent with the exposure preventing the outcome in more than 5% of exposed participants if it also causes the outcome in some other exposed participants (Cartwright, 2010). In neither of these two scenarios does the effect size reveal the frequency with which the exposure caused or prevented the outcome. The effect size simply reveals the net difference in outcome causally attributable to the exposure (Broadbent, 2013, pp. 53–54).

[9] The development of the Framingham scale involved statistical modelling, but the algorithm still extrapolates from outcomes measured in a (cohort) study.

Returning to the generalization inference, a crucial assumption is represented by the ellipsis in the scheme above: a representativeness assumption, stating that the study population is *sufficiently similar* to the target population. With this assumption we are justified in expecting a similar effect, but evidence is required to support it. The evidence might consist in certain methodological features of the study, or in empirical evidence for causal comparability. As we will now see, we should be concerned that we often do not have the needed evidence, that the representativeness assumption fails in typical extrapolations, and that the assumption is poorly articulated in medicine.

4.2. The trouble with the generalization

Having described its structure, we will now discuss the main problem with the Risk GP generalization scheme: trouble with its representativeness assumption. As Broadbent argues (2013, pp. 107–108), the entire work of extrapolation is done by this assumption. Yet several philosophers and medical commentators are worried that in extrapolating—in particular, in extrapolating treatment effects—we often fail to take account of the assumption, even when the representativeness of the study should be in doubt.

Evidence for sufficient similarity is typically one of two kinds: methodological or causal-empirical. Methodological evidence for the representativeness of the study considers how the study population was sampled; in particular, if the study population was a large random sample from the target population, we might assume (perhaps after checking a few variables known to be important) that the effect size statistic in the study represents the effect size statistic in the target. This judgement also demands that the study protocol creates study conditions—including lifestyle circumstances and environmental exposures—that are representative of the conditions the target population will encounter.

Unfortunately, study populations are not always samples from the target population. In extrapolating the effect size measured in a clinical trial, our target might be a population that was ineligible for the trial, such as older patients or patients with other concurrent diseases. If the study population is not a sample from the target population, then it cannot be a *random* sample from the target.

Even when the extrapolation is from a sample to the population sampled—what Rudolf Carnap (1945) called the "inverse inference"—clinical trials, the studies from which we most commonly extrapolate effect size estimates, seldom sample randomly (Bluhm & Borgerson, 2011). Instead, strict inclusion and exclusion criteria are used to determine the wider population eligible for enrolment, and factors such as proximity to the trial site further decide which patients are actually enrolled. Thus, in extrapolating the effect size, methodological reasons for assuming representativeness are rarely satisfied.[10]

In the absence of a methodological warrant for extrapolation, we need some form of causal-empirical evidence. In general, this evidence consists of two parts: knowledge of the variables appearing in the causal mechanisms from exposure to outcome, and empirical data on the distribution of these variables in the study compared to the target. For instance, assuming that the

[10] There have been recent calls for conducting 'pragmatic trials' (vs. 'explanatory' or 'efficacy trials') in medicine, embodied in a series of articles published in the *Journal of Clinical Epidemiology* (Zwarenstein & Treweek, 2009). Pragmatic trials are designed to inform treatment decisions by deliberately enrolling typical patients in typical care settings. They are thus ideal candidates for studies that provide methodological evidence for representativeness.

probability of an effect is fixed by its causes, Nancy Cartwright articulates sufficient conditions for extrapolating the effect size from study population X to target population θ: (a) X and θ are the same with respect to "[t]he causal laws affecting O", and (b) "[e]ach 'causally homogeneous' subclass has the same probability in θ as in X" (2011, pp. 754). The latter condition (b) specifies that the causal variables are distributed identically in the study and the target, while the former condition (a) ensures that if the causal variables are distributed identically, their contribution to the outcome will be the same. Cartwright emphasizes how epistemically demanding these conditions are. She argues that this depth of knowledge is never obtained in reality, but that moreover, we should not expect any set of sufficient conditions for extrapolation to hold widely, so the effect size will seldom be transportable.

Similarly, Daniel Steel argues that "similarity in all relevant respects may be required for extrapolating an exact, quantitative causal effect" from a study or "base population" to a target population (2008, pp. 80).[11] Like Cartwright, Steel concludes that because background causes always differ between two populations, quantitative causal effects will rarely be replicated in a target population, even if the target is closely ancestrally related to the study population.

We present the proposals of Cartwright and Steel only to illustrate the kinds of causal-empirical knowledge that could plausibly allow us to extrapolate the effect size. It is possible to conceive of other conditions for generalizing. However, if these other conditions are anywhere near as demanding, we must concede that the kinds of knowledge and data needed will be difficult to come by.

Several authors in the medical literature have also presented considerations for extrapolation that attempt to locate causally relevant similarities and differences between two populations (Cowan & Wittes, 1994; Dans, Dans, Guyatt, & Richardson, 1998; Dekkers, von Elm, Algra, Romijn, & Vandenbroucke, 2010; Post et al., 2013; Rubins, 1994). Unfortunately, these suggestions typically lack rigour, and fall far short of spelling out the conditions needed to validate the generalization inference. For instance, under the heading of "Process of evaluation of compelling reasons that might limit generalizability", Post and colleagues present a list of six questions adapted from the "User's Guides" to assessing the applicability of clinical trial results (Dans et al., 1998). The questions include: "Are there [biological] patient differences that may diminish the treatment response?", and "Are there important differences in patient compliance that may diminish the treatment response?" (Post et al., 2013, pp. 642). Their approach to assessing group comparability is literally question begging; it provides us with no answers as to when patient differences will modify the effect size. Thus, the standard model of generalization is *too* elliptical, incomplete as a model for medical prediction.[12]

The claim made by Cartwright and Steel that effect sizes are seldom transportable to different populations is an empirical one; its truth depends on how often in reality our study and target populations are causally comparable in all of the right ways. Many medical commentators seem to share their concern; they worry that our randomized clinical trials in practice fail to represent our target populations. Thus, the "external validity" or "generalizability" of these trials—the extent to which the trial results predict the results of intervening elsewhere—might be poor (Black, 1998; Campbell-Scherer, 2010; Feinstein & Horwitz, 1997; Rawlins, 2008; Rothwell, 2005; Upshur, 2005).

Concerns about the generalizability of our evidence base are legitimate. Not only do investigators seldom aim for representativeness in clinical trial design, but the routine design of these trials promotes gross *unrepresentativeness*. For instance, enrolment criteria in trials often exclude older patients, patients with multiple diseases, and patients taking multiple medications (Van Spall, Toren, Kiss, & Fowler, 2007). Yet given the demographics of patients in hospital and community practice (Bajcar et al., 2010; Goulding, Rogers, & Smith, 2003; Salisbury, Johnson, Purdy, Valderas, & Montgomery, 2011), target populations often include these very patients that trials exclude. Age-related physiological changes, diseases and medications might causally interact with the treatment, modifying the effect size. Thus, we should in the very least be cautious about extrapolating effect sizes from typical trials to typical target populations. Since the conditions for sound extrapolation are poorly articulated in medicine, we might worry that practitioners reasoning within the Risk GP generalization scheme would default to the conclusion without careful attention to the causal context of study and target.[13] In these cases, the assumption of representativeness—however we choose to formulate it—may fail, and we may be led into error.

5. Particularization

5.1. The inference scheme

In target population F, AR = r or ES = r

...

For patient a in target population F, $p(O|F) = r$ or $\Delta p(O|F) = r$

Generalization is a predictive inference in which we project either the frequency of O or the effect size from a study population onto a target population. The term 'generalization' might be preferred by some authors to emphasize that it is an inference from the less general (a relatively small population, often atypical with respect to its clinical features) to the more general (usually a much larger, more typical patient population). For the sake of contrast, we call the subsequent inference a 'particularization' because it is an inference from the less particular (a population) to the more particular (a patient).

The same clinical sources that discuss generalizing or extrapolating often do not describe the subsequent inference from group aggregate outcomes to individual patient outcomes. The inference is suppressed—but some prediction activity must be at work if medicine is to have anything to say about a patient's prognosis or about the potential benefits and harms of treatment for that patient. If we do indeed make inferences about patients in clinical medicine, then what might those inferences look like?

We will first consider whether the standard approach in prognosis is to infer the full-stop claim that 'patient a will O'. Carnap considered this type of induction a special case of the "direct inference"—the inference from a population to a sample drawn from the population—in which the sample has an n of 1 (1945, pp. 84–85). If the absolute risk of O is less than 100% but still high in target population F (most Fs will O), we might predict that *this* F will O. We can construct the n-of-1 direct inference as follows.

[11] Steel also defines a less restrictive sense of representativeness in which the base and target are "cell representative" (2008), pp. 205–208, however the conditions under which we can extrapolate according to the cell representativeness criterion are only slightly less demanding.

[12] Cartwright (2007) forcefully criticizes our effectiveness predictions for their lack of rigour.

[13] In evidence-based medicine, the conclusion that trial results are generalizable is indeed treated as a default, and one is instructed to look not for evidence of representativeness but for evidence that the present case is an exception to the rule. For an example of this approach, see Post et al. (2013); for a critique of their position, see Fuller (2013b).

Most *Fs* will *O*
a is an *F*
a will *O*

When the outcome is very common, physicians may indeed reason according to this scheme. One might worry about the correctness of the inference. For example, it does not preclude the possibility that *a* is also a *G*, and that most *G*s will ¬*O* (an instance of the reference class problem, discussed further in Section 5.2). Yet even if the inference were unproblematic, the descriptive accuracy of the model for medical prediction is certainly *not* unproblematic. Medicine often deals with outcomes that are not very common but are still important—either they are highly valued or highly disvalued. The Framingham Risk Scale considers as a high-risk group (*F*s) a population in which the frequency of a CVD event (*O*) is 20%. Despite the fact that most *F*s in this group will not develop CVD, a physician will often act to prevent CVD by prescribing a statin medication. If we assume that physicians have *some* justification (however sound or unsound) for treating high-risk patients with statins, we must search for the justification beyond the *n*-of-1 direct inference scheme.

Instead of a full-stop, definite prediction claim, physicians are instructed to communicate the patient's risk of the outcome (Goodman, 1999). The slip from talking about risk in a population to talking about a patient's risk is subtle but not trivial; it signals a shift in meaning from 'risk' as frequency of the outcome to 'risk' as probability of the outcome. The Framingham Risk Scale does allow us to predict the future prevalence of CVD in an untreated high-risk population (risk as frequency). But the Risk Scale is primarily intended as a tool for bedside medicine, so it seems unlikely that the ultimate aim of the scale is to predict the frequency of CVD in a population. Rather, the investigators of the Framingham Heart Study note that the purpose of their study was "to formulate a single multivariable risk assessment tool that would enable physicians to identify *high-risk candidates* for any and all initial atherosclerotic CVD events using measurements readily available at the clinic or office" (D'Agostino et al., 2008, pp. 744; emphasis added). The purpose of the Framingham Risk Scale is to estimate a patient's risk of CVD (risk as probability) so that therapy can be rationally considered.[14]

The transition from frequencies to probabilities often goes unnoticed, and thus the two-stage inference involved in standard medical prediction is sometimes presented as a single step (Guyatt et al., 2008, pp. 6). However, particularization is an inference distinct from the generalization from study population to target population. In prognosis under the Risk GP Model, particularization is an inference from the premise that the frequency or absolute risk of the outcome in a target population *F* is *r* to the conclusion that the probability of the outcome for a patient—given that the patient is an *F*—is *r*. Communicating a patient's risk in standard practice means expressing a probability equal to the absolute risk we predict in a reference population within which the patient can be placed (Djulbegovic et al., 2011; Goodman, 1999).[15]

The ellipsis in the particularization scheme above represents further needed assumptions. One indispensable assumption is that the patient under consideration is indeed a member of the reference population: *patient a is an F*. If the patient is not a member of population *F*, then there are no grounds for measuring the probability of the outcome for the patient from the frequency of the outcome in *F*.

A second important assumption is invoked to support the transition from frequencies to probabilities: the probability of each member of target population *F* is the same ($F = \{x_1, x_2, \ldots, x_n\}$, $p(x_1) = p(x_2) = \ldots p(x_n)$). No member of *F* is any more or less likely the patient whose individual risk we are presently estimating.[16] To illustrate this idea, imagine that all members of the target population were registered in a large central database. Imagine also that we had a randomizer—perhaps a computer program—that randomly selected a patient from the list, so that each patient had the same probability of being chosen, equal to $1/n$ (where n is the size of the population). If the absolute risk or frequency of *O* will be 2% and there are 100 patients on the list, we predict that 2 patients will have the outcome. The probability of randomly selecting one of these two patients that will develop the outcome is $p(O|F) = 1/100 + 1/100 = 2/100$, which is equal to the absolute risk. However, if the randomizer was broken and as a result one of these two patients was ten times as likely to be chosen compared to any other patient on the list, then $p(O|F) = 10/100 + 1/100 = 11/100$, which is *not* equal to the absolute risk. The assumption that the probability of each member of the target population is the same is crucial for the particularization inference. In the next section, we will consider the extent to which this assumption and the assumption that patient *a* is an *F* are reasonable in clinical medicine.

Just as physicians are often interested in uncommon but meaningful outcomes in prognosis, in treatment they are often interested in small but clinically important effect sizes. Recall that the CVD risk difference attributable to a statin is 5% in a target population with an untreated risk of 25% ($RD = AR_{\neg X} - AR_X = 25\% - 20\% = 5\%$). If we had to predict anything definite, we would predict that the patient will not have the outcome either way because the frequencies of *O* among treated and untreated patients—25% and 20%, respectively—are well below 50%. Yet guidelines recommend treating this population with a statin (Genest et al., 2009) because CVD events are so undesirable and statins are thought to decrease the patient's risk.[17] Similar to prognostic particularization, treatment particularization under the Risk GP Model is not a predictive inference to outcomes that we predict will occur but an inference to probabilities of occurrence. The same two assumptions needed in prognosis are thus also needed in treatment: the patient is a member of the reference population, and the reference population is an unbiased probability-generating setup.

In treatment, particularization involves the effect size. We saw in the previous section that interpreting the effect size is tricky. It might be an error to believe that the risk difference reveals the frequency with which the exposure will cause the outcome in the target

[14] Many authoritative sources use the concept of risk as probability. Paul Glasziou and Les Irwig write, "To identify patients who should expect benefit to be greater than harm, we need to predict each patient's risk" (1995), pp. 1358. The *Users' Guides* claim, "Clinicians require studies of prognosis— those examining the possible outcomes of a disease and the probability with which they can be expected to occur" (Guyatt et al., 2008), pp. 511. In the Preface to a *Lancet* volume called *Treating Individuals*, Peter Rothwell raises a key question that he claims is frequently asked by clinicians applying study results: "How can I judge whether the probability of benefit from treatment in my current patient is likely to differ substantially from the average probability of benefit reported in the relevant trial or systematic review?" (2007), pp. ix.

[15] We should emphasize that the probability generated by Risk GP is a conditional probability: $p(O/F)$. The $p(O/F)$ is a probability 'for a patient' insofar as that patient is an *F*. It should not be confused with some single case probability or chance that is fixed by all of the physical facts relevant for that patient. If such a non-trivial single case probability exists, its value will most likely diverge from the $p(O/F)$ because the target population *F* is typically heterogeneous.

[16] To use Ian Hacking's terminology, the chance setup—for us the target population along with the mechanism for selecting patients for risk assessment—must be *unbiased* (Hacking, 2001), pp. 24–25.

[17] Prudential judgements—those involving a joint consideration of the probability and desirability of the outcome—are essentially judgements of the expected utility of a course of action, even when the reasoning is qualitative or not made fully explicit.

population. It might thus be equally erroneous to infer the probability that the treatment will cause the outcome from the risk difference.

Instead, we infer the *change* in probability of the outcome due to the exposure ($\Delta p(O|F) = r$) from the effect size attributable to the exposure (ES = r). If the effect size is measured as the relative risk, then the relevant change in probability is $p(O|X\&F)/p(O|\neg X\&F)$ because $RR = AR_X/AR_{\neg X} = p(O|X\&F)/p(O|\neg X\&F)$. However, if the effect size is measured as the risk difference, then the relevant change in probability is $p(O|\neg X\&F) - p(O|X\&F)$ because $RD = AR_{\neg X} - AR_X = p(O|\neg X\&F) - p(O|X\&F)$. In deciding whether or not to treat, the ideal comparison modelled by Risk GP is between the probability of O given treatment (for O) and the probability of O given no treatment (for O).[18]

In deciding whether or not to treat, it is important that the change in probability of the outcome reflects the effect size of the treatment rather than a mere correlation between treatment and outcome. Switching your mailing address on a heart health survey from your home to your workplace address might switch your reference population for the study. If your new estimated probability of CVD qua member of the new reference population was different, you would not conclude that changing your address on the form had any effect on your cardiovascular health. Analogously, a physician should not recommend a potentially harmful treatment if all it will do with respect to the desired outcome is shuffle the patient around into a different reference class. They should recommend treatment to a patient only if it is reasonable to believe that the treatment might make a positive causal difference in the outcome for that patient.

An error in our upstream causal inference (e.g. concluding that there is an effect size due to mailing address from a confounded correlation between CVD and mailing address) can lead to an error in our downstream decision-making. More generally, an error in *any* of our upstream inferences, including the generalization or the particularization, might infect our ultimate conclusions. An inferential chain is broken if there is a chink in any one of its constitutive links. The particularization scheme is certainly not without its weaknesses. As with generalization, there are problems with its core assumptions, which we will now describe. We will then argue that standard particularization fails to address a further issue that probabilistic inferences involving single cases (e.g. patient a) must confront: the reference class problem.

5.2. The trouble with the particularization

In contrast with generalization, particularization is less often discussed in the medical literature; it is largely an implicit step in our medical prediction activities. Yet probabilistic reasoning in general *has* received serious attention in the philosophy of science, and as compared to generalization, the two assumptions needed for valid particularization are easier to formulate.

First, we must assume that patient a is an F, or else the probability of the outcome in population F is not directly relevant for a. Though the assumption might seem trivial from the perspective of the inference scheme (it is presupposed by the conclusion), it is non-trivial from the perspective of medical practice. Careful work must be done to establish that the assumption is true. If 'F' reflects some parameter that we can ascertain with relative ease such as a high-risk score on the Framingham Risk Scale, we might be fairly certain that the patient is an F. However, if 'F' reflects a diagnosis for which we do not have decisive evidence, then our certainty in the truth of the assumption will be significantly less. Our uncertainty as to whether the patient is an F should then influence our uncertainty as to whether the patient will have the outcome.

For the sake of illustration, say that you have an urn containing spotted or freckled balls (F), as well as balls that are not freckled ($\neg F$). Additionally, say that half of the freckled balls are orange (O). Meanwhile, none of the not-freckled balls are orange. So $p(O|F) = 0.5$, while $p(O|\neg F) = 0$. Finally, precisely half of all of the balls in the urn are freckled, $p(F) = 0.5$. What is the probability that you will randomly draw an orange ball from the urn, $p(O)$? It would be arbitrary (and mistaken) to set your probability of O to $p(O|F)$ because we cannot assume that the ball you draw will be freckled; it is just as likely that the ball will be not-freckled. The correct probability is the probability of O given a ball of any colour. Since a quarter of all balls in the urn are orange, $p(O) = 0.25$ ($p(O) = p(O|F)p(F) + p(O|\neg F)p(\neg F) = (0.5)(0.5) + 0(0.5) = 0.25$). From the equation for $p(O)$ just used, we see that it will only be the case that $p(O) = p(O|F)$ when: (i) $p(F) = 1.0$ (we are certain that the ball will be F), or (ii) $p(O|F) = p(O|\neg F)$ (the probability of O is the same regardless of whether or not the ball is F). Otherwise, how closely $p(O|F)$ approximates $p(O)$ depends on how close is our certainty to 1.0 that the ball will be F, and how close the $p(O|\neg F)$ is to the $p(O|F)$.

In short, if we lack certainty about an individual's membership in F, we might not want to assume it for the sake of inference. Since we are often relatively uncertain about a patient's diagnosis or about their membership in a risk group given fallible evidence, it would often be imprudent to set our probability of O to $p(O|F)$. The Risk GP particularization does not accommodate diagnostic or classificatory uncertainty, and provides no direction when confronted with this frequent phenomenon. When both $p(O|F)$ and $p(O|\neg F)$ can be estimated, it would be better to reason according to the equation for $p(O)$ (total probability) above.

Particularization also relies upon a second assumption that is sufficient for deriving $p(O|F)$ from the frequency of O in target population F: the probability of each member of target population F is the same. Each member of the target population must have an equal probability of having their risk assessed. We modelled this situation with a database that included all patients in F and a computer program that randomly chose a patient from the list.

We might wonder how well this probability model represents our target populations in medicine. We can start by considering the target populations defined by clinical practice guidelines. Guidelines are often produced by national professional groups and are usually disease-specific, so their target populations will often consist of all patients with a certain diagnosis in a particular country. We might consider as a target population all patients on the UK's National Health Service (NHS) with a certain diagnosis. If guidelines are supposed to guide individual clinicians, then the relevant unknown is the probability of the outcome for a patient on this list who is assessed by a given clinician. However, clearly not all patients in the NHS have an equal chance of being seen by the same doctor; those patients who live in the north of Scotland have a very low chance of being seen by a physician in the south of Wales. Thus, the probability model assumed by Risk GP is a poor fit for an obvious example of a target population setup: a clinician applying a practice guideline.

Perhaps instead we should define target population F to include only those patients that are likely to be assessed by a given clinician; for instance, everyone listed in a community physician's regular patient roster. Even then we should doubt that each patient in the target population has an equal probability of being assessed; some patients are rarely seen by their physician at all. So our probability model might also inadequately represent a physician assessing their registered patients. If assessment is correlated with

[18] In deciding between two alternate treatments (X and Z), the comparison is instead between the probability of O given X (and not Z) and the probability of O given Z (and not X).

the outcome, then the probability of the outcome for patients who are assessed will differ from the frequency of the outcome in the overall target population. For example, it is reasonable to suspect that patients with vague, undiagnosed cardiovascular symptoms like occasional shortness of breath or minor chest pain are more likely to have their risk of CVD assessed by their physician using the Framingham scale. Since these symptoms sometimes indicate underlying CVD, assessment might be correlated with CVD, even among Framingham high risk patients (the target population). The assumption that the probability of each patient is the same might not hold in the context of our most natural examples of target population setups, which may lead us to greatly misestimate the probability of the outcome.

Aside from the trouble with its two assumptions, there is one final weakness of the particularization scheme worth visiting, a problem with its conclusion. Most often target population *F* will be a broadly defined group such as patients with a particular disease. It is doubtful that the risk measure calculated in a study would be generalizable to many more narrowly defined target populations as the relevant features according to which we might want to narrow our target population are the same features that would ostensibly modify the value of the risk measure. But when we can confidently predict how a risk measure will differ according to other features particular to the patient, we should. Many authors in the medical literature have decried that our approach based on study averages often loses sight of the relevant particularities of individuals (Feinstein & Horwitz, 1997; Greenhalgh, Howick, & Maskrey, 2014; Kravitz, Duan, & Braslow, 2004; Tanenbaum, 1993; Tonelli, 1998).

The perennial reference class problem looms in the background. As the problem goes, a member of a population is also a member of many subgroups of that population in which the probabilities might differ. For practical purposes, one response makes intuitive sense: we should make decisions based on the probability in the narrowest informative reference class for which we can form a reasonable probability judgement (Flores, 2015).[19] If your high CVD risk patient is a member of the subgroup of high-risk patients that currently report crushing chest pain, then you are well advised to act based on the probability of a heart attack in this more narrowly defined group.

Another way to frame the response is as a matter of more evidence. We might have good evidence for the prediction claim beyond the study from which we extrapolated. This evidence might consist of the results of a different epidemiological study, or non-epidemiological evidence such as the clinician's experience suggesting that patients similar to the one under consideration fair differently on average. Sensibility (in addition to a body of philosophical literature) dictates that we should condition our belief in the hypothesis on all of the available evidence, and not only on the results of a single study, however rigorous that single study may be.[20]

The final problem with the particularization is thus that it is not particular enough; it fails to consider other features beyond 'F' particular to the patient, even when we have good evidence that the probability given 'F' plus these other features diverges from the probability given 'F' alone. Within the Risk GP prediction model, we are—at least sometimes—ignoring information relevant to prediction.

6. Trouble with the standard model and the case for model pluralism

Confirmation that there is indeed a standard model of medical prediction is provided by the fact that there seems to be a common target of several compelling critiques in medicine and in the philosophy of science. There are serious problems with the Risk GP Model, especially with its assumptions, which are often difficult to warrant with evidence and will often fail in practice. These problems are not challenges to the model's validity, but rather to its soundness in many instances of its use, and thus to its privileged status as the standard model.

Since the assumptions are peculiar to the Risk GP Model, so too are the problems with the assumptions. The model is not a constraint on medical prediction because—as we have seen—there are other ways of predicting. Other models of prediction rely on different assumptions, so that another model's assumptions might hold when Risk GP's assumptions fail. Yet the relative lack of attention that other models receive in medicine suggests a certain inflexibility when it comes to prediction.

Medical prediction-makers are best served by adopting a pluralist approach to prediction: exploring all of the options to decide which model best fits the situation at hand. The first step towards model pluralism is to recognize that there are other ways of predicting. To that end, we introduce just a few alternate models of prediction suggested by the preceding discussion. No doubt they each have their own weaknesses. We present them simply to make the case that model pluralism is a practicable alternative to a one-model-fits-all approach.

We saw that two distinct approaches to prediction have coexisted throughout history, one relying on theory or mechanisms, and the other relying directly on experience. In mechanistic reasoning, we start from known causal mechanisms from *X* to *O*. We then predict that intervening with *X* will produce *O* (Andersen, 2012; Howick, 2011b). In so doing, we make several assumptions; for instance, that the mechanisms are understood in their full complexity, that intermediate components of the mechanisms are intact, that the mechanisms produce *O* with a high enough probability, and that the effect of *O*-producing mechanisms is not masked by the presence of *O*-inhibiting mechanisms (Clarke, Gillies, Illari, Russo, & Williamson, 2014; Howick, 2011b).

Because it does not depend on aggregate outcomes in populations, mechanistic reasoning may be useful when *any* of Risk GP's assumptions fail or when we lack human studies altogether. It is most dependable when the relevant mechanisms are simple and well-established. As an example, osteoporosis medications are commonly prescribed to prevent bone fractures (especially hip fractures) in older women. A physician may worry that a particular patient is less likely to benefit from osteoporosis medication than the average target patient, or may have good reason to believe that the patient will benefit but may worry about harmful side effects. In hopes of preventing fractures, they may put aside the empirical evidence on osteoporosis medications and instead reason that osteoporotic bone fractures are almost always caused by falls. They might then suggest sensible measures to prevent falls in the home.

As another example, in medical conditions caused by nutrient or hormone deficiency (e.g. iron deficiency, hypothyroidism), we may reverse the condition very simply by supplying the nutrient or hormone in which the patient is deficient. We rarely have clinical research evidence demonstrating the effectiveness of various doses of replacement for various magnitudes of deficiency. In lieu of extrapolating from research studies, the physician might tailor the

[19] We cannot defend the intuition here, but see discussion in Gillies (2000), pp. 119–123. Early notable 'narrowest reference class' solutions to the reference class problem were given in Keynes (1921) and Ayer (1963). As Brendan Clarke et al. argue (2014), features suggested by the relevant causal mechanisms may help us to locate the relevant narrower reference class.

[20] For instance, see Good (1967) and Ramsey (1990). Jacob Stegenga (2011) criticizes the lack of evidential inclusiveness in medical treatment meta-analyses, studies from which we often extrapolate an effect size estimate.

replacement dose to restore a particular patient's levels to the statistically normal range, especially if the deficiency is asymptomatic (and thus the dose cannot be adjusted based on patient feedback). This mechanistic reasoning is based on the simple principle that if a deficiency causes a condition, curing the deficiency cures the condition.

In contrast to mechanistic reasoning, predicting from personal clinical observations is a model that often relies on an induction from previous experience. There are two contexts in which it may be useful: predicting from experience with previous patients, and predicting from experience with the same patient. Predicting from experience with previous patients may be useful when the research literature lacks data on a relevant patient subgroup into which the present patient falls; for instance, when the relevant subgroup is hard to operationalize. Based on behavioural cues from their patient (the way the patient describes their symptoms, their affect, their non-verbal communication), a physician may experience the intuition that the patient is seriously unwell and at risk for further deterioration in their health (a poor prognosis). It is difficult to systematically study the subgroup 'patients whose behaviour triggers the clinical intuition that they are at risk of deterioration in their health', and an expert physician may have to rely on their own experience with such patients in order to make a prognostic inference.

On the other hand, predicting from experience with the same patient is a useful approach in treatment when the patient's response is observable; for example, in treating symptoms. From previous response to treatment, we predict future response to treatment; or from previous superior response to one treatment compared with another, we predict future superior response. We must assume that the patient's current physiology is sufficiently similar to their previous physiology, but this judgement is often easier to make and more reliable than the assumption that a study population is sufficiently comparable to a target population. The former assumption might be warranted when the latter assumption is not. Even if the former assumption is satisfied, the inference is not infallible; for instance, perhaps the patient's previous symptoms spontaneously remitted soon after starting treatment, and the physician is wrong to infer that the treatment will relieve the patient's symptoms the next time they recur. Thus, the inference is strongest when the patient's untreated symptoms or health status are stable rather than fluctuating. The *n*-of-1 trial can be seen as a refinement of this approach to address its weaknesses (Hankey, 2007). In an *n*-of-1 trial, a patient alternates between treatment and no treatment (or between different treatments), outcomes are systematically recorded, and effectiveness and safety are inferred, sometimes with the aid of statistics.

Other models that maintain a role for controlled studies in predicting the effect of an exposure or intervention but eschew extrapolating the effect size include the proposals of Steel and Cartwright. Although both of their approaches require detailed background knowledge, in contrast to the generalization inference scheme, the conditions for sound prediction are explicitly formulated and thus may be easier to assess. According to Steel (2008), even if we are unable to extrapolate the effect size from a study, we might still be able to extrapolate the "positive causal relevance" of the exposure. After formulating the notion of positive causal relevance in terms of ideal causal Bayes nets interventions, he develops a "mechanisms approach" to extrapolating positive relevance. Oftentimes decision-makers wish to know an exposure's effect size to help decide among alternate courses of action. Other times, predicting that the exposure will have a positive (greater than zero) effect in the target provides enough information. To illustrate, Steel develops his theory mainly in the context of extrapolating from animal studies, especially extrapolating harmful effects like carcinogenicity. We may lack good epidemiologic data but possess data from controlled animal experiments showing that an exposure causes serious harm in animals. If we can extrapolate that the exposure will also cause serious harm in at least *some* humans, this conclusion might warrant public health action; we might disseminate a public advisory warning, regulate the use of the exposure, or ban its use altogether.

Meanwhile, when we have controlled human studies but they are not sufficiently representative of the target, they might serve as a proof of principle that the exposure *can* make a difference rather than a base from which to extrapolate the effect size. Cartwright (2012) advises asking whether the causal principle or mechanism that operated in the controlled study will also operate in the target. Predicting that the intervention "will work for us" also requires knowing that certain support factors are present in the target context to enable the intervention's causal power. Her account sheds special light on predictions involving policy or social interventions. Extrapolating from clinical studies of behavioural or complex care interventions is often difficult because the healthcare context strongly determines effectiveness and will often vary across sites, even within the same health system. An educational intervention to promote diabetes self-management will only work if the patient has the local resources to make lifestyle changes. Similarly, the effectiveness of an interprofessional team intervention aimed at developing comprehensive care plans for patients with multiple chronic conditions might depend on the composition of the care team and the patient's particular constellation of diseases.

Probabilistic models to rival the Risk GP particularization scheme are also available. As argued in Section 5.2, in prognosis when we know the $p(O|F)$ and the $p(O|\neg F)$ but are relatively uncertain as to which group—F vs. $\neg F$—the patient belongs, we can calculate the probability of the outcome according to the formula for total probability: $p(O) = p(O|F)p(F) + p(O|\neg F)p(\neg F)$. This model may be useful when we lack a definitive diagnostic inference that can decide whether the patient is F or $\neg F$—a not uncommon scenario. Diagnostic tests often lack sensitivity (given a negative result we are not confident that the patient lacks the disease) or specificity (given a positive result we are not confident that patient has the disease). But from the negative or positive finding and the reported sensitivity or specificity we can calculate the probability that they have the disease, $p(F)$, and the probability that they do not have the disease, $p(\neg F)$. If we also know the outcome rates among individuals with and without the disease, we can determine the $p(O|F)$ and the $p(O|\neg F)$, respectively. Finally, we can calculate the patient's prognosis, the $p(O)$, from the total probability equation, based on our current diagnostic evidence. We can then determine whether they fall into a useful risk category for treatment purposes.

On the other hand, we saw that the standard model neglects evidence for the probability of the outcome beyond the study population from which it extrapolates. We might know that the patient is an F and the $p(O|F)$, but are well-advised to update our probabilities in the light of further evidence or further features particular to this patient (assuming that the evidence is credible). We could choose to update our probabilities quantitatively and formally, using—for instance—the model of Bayesian conditionalization sometimes used in diagnosis. The Framingham Risk Scale tells us the probability of a CVD event given a set of risk factors 'F', the $p(O|F)$. We might additionally learn that the patient has a family history of CVD in a first-degree relative, 'H', and may wish to revise the probability of a CVD event accordingly, since patients with a family history of CVD are more likely to have CVD themselves. Bayes' Theorem tells us that $p(O|H) = [p(H|O)p(O)]/p(H)$. We can treat the risk of CVD among patients with risk factors 'F' as the base rate or prior probability of CVD, $p(O)$. We can then use

other epidemiologic evidence to determine the probability of having a family history of CVD among patients with CVD, p($H|O$), as well as the probability of having family history of CVD in general, p(H). Finally, we can calculate the probability that the patient will have CVD given the new information that they have relevant family history, the p($O|H$), using Bayes' Theorem.[21]

Other times, probabilities may be difficult to estimate; and besides, physicians and patients might struggle to make meaning out of the precise numbers. Thus, we might instead choose to update our probabilities informally, or even qualitatively. We might expect a patient's magnitude of benefit from a treatment to be lower if they are poor adherer to the treatment. We may not know exactly how much lower it will be, but depending on our purposes, it might still be useful to expect a lower risk reduction than inferred by Risk GP. Among patients that achieve some reduction in LDL cholesterol using statin therapy but fall short of evidence-based targets, we might expect the relative risk to be higher than the 0.8 predicted from trial evidence (less of a risk reduction). Thus, the reduction in probability of CVD due to statin therapy will also be lower. Among patients that achieve *half* of the recommended reduction in LDL cholesterol, we might predict that the relative risk will be roughly 0.9 (half as many CVD events prevented). For the patient with an untreated risk of 25%, this prediction suggests a reduction in probability of CVD from 25% to 22.5%, compared to the reduction from 25% to 20% predicted by Risk GP.

Considering that alternate models are possible, we should question whether we need a standard model of prediction in medicine, a strategy that seems to promote the inflexibility of reasoning we caution against. Medicine would better accommodate the context-sensitivity of its prediction activities if it were instead standard to carry along a plurality of models and to match the model to the circumstances. Unfortunately, authoritative medical textbooks and guides that teach the gold standard of prediction fail to mention many of the alternatives. Other models, such as mechanistic reasoning and predicting from personal experience, receive variable recognition and often scant exposition. Crucially, because the standard model is elliptical—its assumptions remain largely unarticulated in standard practice—we are often unable to recognize when the model fails, are often unable to appreciate the importance of alternate models, and often fail to take model pluralism seriously. Yet when it comes to medical prediction, many models are surely better than one.

7. Conclusion

'Prediction' refers to prediction activities as well as prediction claims, broadly construed to cover unknown events and outcomes, or narrowly construed to cover only future events and outcomes. In medicine, narrow predictions are made in prognosis and treatment, while broad predictions are made in diagnosis. Induction from experience is one rational approach to prediction used throughout medical history from at least the time of Hippocrates. It matured with the ascent of epidemiology and statistics, giving rise to the Risk Generalization–Particularization Model, the standard approach to prediction in contemporary medicine. The generalization involves extrapolating a risk measure (either the absolute risk or the effect size) from a study population to a target population, while the particularization involves probabilizing the risk measure for a patient from the target population.

The standard model is not without its troubles. Risk GP models medical prediction incompletely; further assumptions are needed in place of the ellipses. First, generalization requires a representativeness or sufficient similarity assumption. Unfortunately, in treatment the effect size may not generalize widely due to our unrepresentative efficacy studies. Moreover, the conditions for sound extrapolation are poorly articulated in standard practice, resulting in an overly elliptical generalization scheme. Meanwhile, particularization depends upon two assumptions. The first assumption, that the patient under consideration is a member of the reference population, is questionable whenever there is significant diagnostic uncertainty. The second assumption, that the probability of being assessed is the same for each patient, is tenuous in the case of perhaps our most natural target populations. Finally, standard particularization is frequently not particular enough, failing to consider further features that locate the patient in a narrower informative reference class. There are of course other models of prediction we could use. Rather than inflexibly committing ourselves to one standard model, it should be standard to embrace model pluralism in medical prediction.

Acknowledgements

Our sincere thanks to Margherita Benzi, Nancy Cartwright, Donald Gillies, Maël Lemoine, David Papineau, Jacob Stegenga, two anonymous reviewers, and audiences at the Prediction in Epidemiology and Healthcare workshop at King's College London, as well as the Philosophy of Medicine Seminar at IHPST, Université Paris 1 for helpful comments and discussion. JF is grateful for support from the Canadian Institutes of Health Research and the W. Garfield Weston Foundation. LJF gratefully acknowledges support from the Chilean National Commission for Scientific and Technological Research (CONYCIT). We have no financial interests to disclose.

References

Andersen, H. (2012). Mechanisms: What are they evidence for in evidence-based medicine? *Journal of Evaluation in Clinical Practice, 18*, 992–999.
Ayer, A. J. (1963). Two notes on probability. In *The concept of a person and other essays* (pp. 188–208). Macmillan.
Baigent, C., Keech, A., Kearney, P. M., Blackwell, L., Buck, G., Pollicino, C., et al., , Collaborators, C. T. T. (2005). Efficacy and safety of cholesterol-lowering treatment: Prospective meta-analysis of data from 90,056 participants in 14 randomised trials of statins. *Lancet, 366*, 1267–1278.
Bajcar, J. M., Wang, L., Moineddin, R., Nie, J. X., Tracy, C., & Upshur, R. E. G. (2010). From pharmaco-therapy to pharmaco-prevention: Trends in prescribing to older adults in Ontario, Canada, 1997–2006. *BMC Family Practice, 11*, 75.
Black, D. (1998). The limitations of evidence. *Perspectives in Biology and Medicine, 42*, 1–7.
Bluhm, R., & Borgerson, K. (2011). Evidence-based medicine. In D. M. Gabbay, P. Thagard, & J. Woods (Eds.), *Philosophy of medicine: Vol. 16. Handbook of the philosophy of science* (pp. 203–238). Amsterdam: Elsevier.
Broadbent, A. (2013). *Philosophy of epidemiology*. Basingstoke: Palgrave Macmillan.
Campbell-Scherer, D. (2010). Multimorbidity: A challenge for evidence-based medicine. *Evidence-based Medicine, 15*, 165–166.
Carnap, R. (1945). On inductive logic. *Philosophy of Science, 12*, 72–97.
Cartwright, N. (2007). Are RCTs the gold standard? *BioSocieties, 2*, 11–20.
Cartwright, N. (2010). What are randomised controlled trials good for? *Philosophical Studies, 147*, 59–70.
Cartwright, N. (2011). Predicting 'It will work for us': (Way) beyond statistics. In F. R. Phyllis McKay Illari, & Jon Williamson (Eds.), *Causality in the sciences*. Oxford Scholarship Online.
Cartwright, N. (2012). Will this policy work for you? Predicting effectiveness better: How philosophy helps. *Philosophy of Science, 79*, 973–989.
Clarke, B., Gillies, D., Illari, P. M., Russo, F., & Williamson, J. (2014). Mechanisms and the evidence hierarchy. *Topoi, 33*, 339–360.
Cowan, C. D., & Wittes, J. (1994). Intercept studies, clinical trials, and cluster experiments: To whom can we extrapolate? *Controlled Clinical Trials, 15*, 24–29.
D'Agostino, R. B., Vasan, R. S., Pencina, M. J., Wolf, P. A., Cobain, M., Massaro, J. M., et al. (2008). General cardiovascular risk profile for use in primary care – The Framingham Heart Study. *Circulation, 117*, 743–753.
Dans, A. L., Dans, L. F., Guyatt, G. H., & Richardson, S. (1998). Users' guides to the medical literature: XIV. How to decide on the applicability of clinical trial

[21] Guidelines for CVD risk assessment instruct physicians to double the risk computed by the Framingham scale when CVD is present in a first-degree relative younger than age 60 (Genest et al., 2009), which is good advice if the p($H|O$)/p(H) is roughly equal to 2.

results to your patient. Evidence-Based Medicine Working Group. *JAMA, 279*, 545–549.
Dekkers, O. M., von Elm, E., Algra, A., Romijn, J. A., & Vandenbroucke, J. P. (2010). How to assess the external validity of therapeutic trials: A conceptual approach. *International Journal of Epidemiology, 39*, 89–94.
Djulbegovic, B., Guyatt, G. H., & Ashcroft, R. E. (2009). Epistemologic inquiries in evidence-based medicine. *Cancer Control, 16*, 158–216.
Djulbegovic, B., Hozo, I., & Greenland, S. (2011). Uncertainty in clinical medicine. In D. M. Gabbay, P. Thagard, & J. Woods (Eds.), *Philosophy of medicine: Vol. 16. Handbook of the philosophy of science* (pp. 299–356). Amsterdam: Elsevier.
Duffin, J. (2010). *History of medicine: A scandalously short introduction*. Toronto: University of Toronto Press.
EBM Working Group. (1992). Evidence-based medicine: A new approach to teaching the practice of medicine. *Journal of the American Medical Association, 268*, 2420–2425.
Feinstein, A. R., & Horwitz, R. I. (1997). Problems in the "evidence" of "evidence-based medicine". *American Journal of Medicine, 103*, 529–535.
Flores, L. J. (2015). Therapeutic inferences for individual patients. *Journal of Evaluation in Clinical Practice, 21*, 440–447.
Fuller, J. (2013a). Rhetoric and argumentation: How clinical practice guidelines think. *Journal of Evaluation in Clinical Practice, 19*, 433–441.
Fuller, J. (2013b). Rationality and the generalization of randomized controlled trial evidence. *Journal of Evaluation in Clinical Practice, 19*, 644–647.
Fuller, J., Sankar, A., & Upshur, R. E. G. (2013). Concepts in evidence-informed medical practice. In J. Hall, K. Piggott, M. Vojvodic, & K. Zaslavsky (Eds.), *The essentials of clinical examination handbook* (7th ed.). (pp. 571–583) Toronto: Thieme.
Genest, J., McPherson, R., Frohlich, J., Anderson, T., Campbell, N., Carpentier, A., et al. (2009). 2009 Canadian Cardiovascular Society/Canadian guidelines for the diagnosis and treatment of dyslipidemia and prevention of cardiovascular disease in the adult-2009 recommendations. *Canadian Journal of Cardiology, 25*, 567–579.
Gillies, D. (2000). *Philosophical theories of probability*. London, New York: Routledge.
Gillies, D. (2005). Hempelian and Kuhnian approaches in the philosophy of medicine: The Semmelweis case. *Studies in History and Philosophy of Biological and Biomedical Sciences, 36*, 159–181.
Glasziou, P. P., & Irwig, L. M. (1995). An evidence based approach to individualising treatment. *British Medical Journal, 311*, 1356–1359.
Glasziou, P., & Mant, D. (2007). Applying results to treatment decisions in primary care. In P. M. Rothwell (Ed.), *Treating individuals: From randomized trials to personalised medicine* (pp. 83–95). The Lancet.
Good, I. J. (1967). On the principle of total evidence. *British Journal for the Philosophy of Science, 17*, 319–321.
Goodman, S. N. (1999). Probability at the bedside: The knowing of chances or the chances of knowing? *Annals of Internal Medicine, 130*, 604–606.
Goulding, M. R., Rogers, M. E., & Smith, S. M. (2003). Public health and aging: Trends in aging – United States and worldwide. *Journal of the American Medical Association, 289*, 1371–1373.
Greenhalgh, T., Howick, J., & Maskrey, N. (2014). Evidence based medicine: A movement in crisis? *BMJ, 348*, g3725.
Guyatt, G., Rennie, D., Meade, M. O., & Cook, D. J. (2008). *Users' guides to the medical literature: Essentials of evidence-based clinical practice* (2nd ed.). New York: McGraw-Hill Medical.
Hacking, I. (2001). *An introduction to probability and inductive logic*. Cambridge: Cambridge University Press.
Hankey, G. J. (2007). Are n-of-1 trials of any practical value to clinicians and researchers? In P. M. Rothwell (Ed.), *Treating individuals: From randomized trials to personalised medicine* (pp. 231–246). The Lancet.
Hempel, C. G. (1962). Deductive-nomological vs. statistical explanation. In H. Feigl (Ed.), *Minnesota studies in the philosophy of science* (Vol. 3, pp. 98–169). Minneapolis.
Hempel, C. G., & Oppenheim, P. (1948). Studies in the logic of explanation. *Philosophy of Science, 15*, 135–175.
Horton, R. (2000). Common sense and figures: The rhetoric of validity in medicine. Bradford Hill Memorial Lecture 1999. *Statistics in Medicine, 19*, 3149–3164.
Howick, J. (2011a). *The philosophy of evidence-based medicine*. Oxford: Wiley-Blackwell.
Howick, J. (2011b). Exposing the vanities-and a qualified defense-of mechanistic reasoning in health care decision making. *Philosophy of Science, 78*, 926–940.
Keynes, J. M. (1921). *A treatise on probability*. Macmillan.
Kravitz, R. L., Duan, N., & Braslow, J. (2004). Evidence-based medicine, heterogeneity of treatment effects, and the trouble with averages. *Milbank Quarterly, 82*, 661–687.
Mathers, C. D., & Loncar, D. (2006). Projections of global mortality and burden of disease from 2002 to 2030. *PloS Medicine, 3*, e442.
Papaioannou, A., Morin, S., Cheung, A. M., Atkinson, S., Brown, J. P., Feldman, S., et al. (2010). 2010 clinical practice guidelines for the diagnosis and management of osteoporosis in Canada: Summary. *Canadian Medical Association Journal, 182*, 1864–1873.
Peltzman, S. (1973). An evaluation of consumer protection legislation: The 1962 drug amendments. *Journal of Political Economy, 81*, 1049–1091.
Post, P. N., de Beer, H., & Guyatt, G. H. (2013). How to generalize efficacy results of randomized trials: Recommendations based on a systematic review of possible approaches. *Journal of Evaluation in Clinical Practice, 19*, 638–643.
Potter, P. (1988). *Short handbook of hippocratic medicine*. Quebec: Éditions du Sphinx.
Rabi, D. M., Daskalopoulou, S. S., Padwal, R. S., et al. (2011). The 2011 Canadian hypertension education program recommendations for the management of hypertension: Blood pressure measurement, diagnosis, assessment of risk, and therapy. *Canadian Journal of Cardiology, 27*, 415–433.
Ramsey, F. P. (1990). Weight or the value of knowledge. *British Journal for the Philosophy of Science, 41*, 1–4.
Rawlins, M. (2008). De testimonio: On the evidence for decisions about the use of therapeutic interventions. *Lancet, 372*, 2152–2161.
Rothwell, P. M. (2005). External validity of randomised controlled trials: "to whom do the results of this trial apply?". *Lancet, 365*, 82–93.
Rothwell, P. M. (2007). Preface. In P. M. Rothwell (Ed.), *Treating individuals: From randomized trials to personalised medicine* (pp. ix–xii). The Lancet.
Rubins, H. B. (1994). From clinical trials to clinical practice: Generalizing from participant to patient. *Controlled Clinical Trials, 15*, 7–10.
Salisbury, C., Johnson, L., Purdy, S., Valderas, J. M., & Montgomery, A. A. (2011). Epidemiology and impact of multimorbidity in primary care: A retrospective cohort study. *British Journal of General Practice, 61*, e12–21.
Spirtes, P., Glymour, C., & Scheines, R. (2000). *Causation, prediction, and search* (2nd ed.). Cambridge, MA: MIT Press.
Steel, D. P. (2008). *Across the boundaries: Extrapolation in biology and social science*. Oxford: Oxford University Press.
Stegenga, J. (2011). Is meta-analysis the platinum standard of evidence? *Studies in History and Philosophy of Biological and Biomedical Sciences, 42*, 497–507.
Straus, S. E., Glasziou, P., Richardson, W. S., & Haynes, R. B. (2011). *Evidence-based medicine: How to practice and teach it* (4th ed.). New York: Churchill-Livingstone.
Szklo, M., & Nieto, F. J. (2007). *Epidemiology: Beyond the basics* (2nd ed.). Boston: Jones and Bartlett Publishers.
Tanenbaum S. J. (1993). What physicians know. *New England Journal of Medicine, 329*, 1268–1271.
Tonelli, M. R. (1998). The philosophical limits of evidence-based medicine. *Academic Medicine, 73*, 1234–1240.
Upshur, R. E. G. (2005). Looking for rules in a world of exceptions: Reflections on evidence-based practice. *Perspectives in Biology and Medicine, 48*, 477–489.
Upshur, R. E. G. (2014). Do clinical guidelines still make sense? No. *Annals of Family Medicine, 12*, 202–203.
Van Spall, H. G. C., Toren, A., Kiss, A., & Fowler, R. A. (2007). Eligibility criteria of randomized controlled trials published in high-impact general medical journals: A systematic sampling review. *Journal of the American Medical Association, 297*, 1233–1240.
WHO. (2011). *Global status report on noncommunicable diseases 2010*. Geneva: World Health Organization.
Zwarenstein, M., & Treweek, S. (2009). What kind of randomized trials do we need? *Journal of Clinical Epidemiology, 62*, 461–463.