SCIENCE
AND PROVEN
EXPERIENCE
JOHANNES

SCIENCE AND PROVEN EXPERIENCE JOHANNES

ISBN 978-91-983575-5-4

Graphic design Johan Laserna
Printed by Media-Tryck, Lund University, Lund 2018

Content

Preface

Johannes, as far as we know you have no motto, no simple phrase that epitomises your ideals and aspirations as an individual, a philosopher, a scientist. These days, few of us have such maxims. Too bad, we say. So let us give you a motto. *Art and Science*.

Art and Science. Why pair these two grand nouns? And why the conjunction? Why not Art *or* Science, or a more complex connective such as *and/or-but-not-in-conflict-with-either*, with its cue that art and science are sometimes, but not necessarily combatants, and that shared territory may or may not exist? The first question is easy to answer. You are a truly clever photographer, a photographer with the eye of an artist. Your photos, like all good photos, capture more than the naked eye sees.

You are also a philosopher who has taken on some of the most difficult, eternal philosophical questions. Doing so, you have given us new insights into old problems. But you are definitely not a philosopher who looks at scientific problems from a distance, away from the practical concerns of science. To you, scientific evidence is important, and so is

proven experience. This is shown in your work with scientists from many different areas – from forestry, nursing and cognitive science, and from jurisprudence, climate research, ecology and medicine.

At this point a few readers, versed in history, will be growing mildly indignant. The motto we have given you is already taken, they will say, and worse, we are wilfully misconstruing its true meaning. For many, many years the motto *Science and Art* has adorned The Royal Institute of Technology in Stockholm. The musings above do little, if any, justice to that history. They don't expound or explain it – and it doesn't help at all throwing the words around, in whatever order! So, what *are* we dealing with here? The problem is not the concept of science. "Science" simply means something like the systematised search for new knowledge. The real problem is that in this context "Art" does not refer to *the arts* – to photography, music, theatre, film, dance...and the like. It means *knowledge how*, as in "the art of conversation". It indicates know-how, the ability to construct or make things on the basis of solid scientific knowledge – things such as a bridge, a molecule, or an autonomous robot. Making things gives us proven experience, and proven experience helps us make things. Just think of making a photograph. Or think about how much proven experience is hidden in the word "konst", and in terms like "metabolisk ingenjörskonst", "krigskonst" and "skeppsbyggarkonst".

So maybe, Johannes, your motto should be, after all, not Art and Science, or Science and Art, but Science and Proven Experience.

By coincidence, it turns out that this motto is also the title of our research programme, the research programme of which you are a cherished member: Vetenskap och beprövad erfarenhet (VBE).

This is the sixth booklet in our series of VBE volumes. It contains twelve essays on science and proven experience. It is our gift to you on your fiftieth birthday. Happy birthday! From all of us.

Collectively, the papers summarise what we have done up to now. Individually, each points to a future – to wonderful, intriguing research questions and problems that lie ahead of us, to questions and problems we are looking forward to discussing with you. Who knows, there might be more than one research paper or a new book hidden in this volume.

Nils-Eric on behalf of the VBE program*

* VBE, Vetenskap och Beprövad Erfarenhet (Science and Proven Experience) was established on 1 January, 2015. VBE is an international and multidisciplinary research programme sponsored by Riksbankens Jubileumsfond (The Swedish Foundation for Humanities and Social Sciences). The programme's researchers represent Lund University, Malmö University and Statens beredning för medicinsk och social utvärdering, Stockholm (the Swedish Agency for Health Technology Assessment and Assessment of Social Services, SBU) in Sweden; Carnegie Mellon University and Harvard Medical School in the US;

and Leeds University in the UK. The programme brings together research in disciplines including philosophy, psychology, cognitive science, jurisprudence, medicine and business.

Information about the VBE-program can be found at vbe.lu.se.

A note on randomized controlled trials in evidence-based medicine

STEN ANTTILA

Worrall (2002) is critical of the heavy epistemic weight given to randomized controlled trials (RCTs) in evidence formation in medicine. He criticizes some arguments in favor of randomization commonly made by the proponents of evidence-based medicine. I will discuss Worrall's criticism of one of these arguments. I call it "the de-confounding argument". My point of departure is the frequentist statistical inference that presently dominates medical research.

De-confounding

The de-confounding argument, according to Worrall (2002), is that all confounders (known and unknown) can be controlled for with randomization (p. 322). This claim, taken literally, is "trivially unsustainable" according to Worrall:

> It is perfectly possible that a properly applied random process might "by chance" produce a division between control and

> experimental groups that is significantly skewed with respect to some uncontrolled prognostic factor that in fact plays a role in therapeutic outcome. (p. 322)

Furthermore...

> The control and experimental groups could be deliberately matched relative to some features, and, despite the qualms of some avid randomizers, surely ought to be matched with respect to factors that there is some good reason to think may play a role in recovery from, or amelioration of the symptoms of, the condition at issue. (p. 322)

He concludes that the de-confounding argument is delusive (p. 328).

Unbiased estimators

I think that Worrall is correct, in that random allocation does not necessarily control for all possible confounders. But there are strong arguments for the de-confounding argument within the very framework of frequentist statistics, and Worrall does not really acknowledge these. He contends (p. 320–1) that the argument "that the logic of the classical statistical significance test requires randomization" is not convincing "even on its own terms", but he declines to elaborate on this. He calls this the Fisherian argument.

The main argument is something like the Fisherian argument, and is about substantiating an assumption that is

required for frequentist inferences to be valid. When an estimator is biased, neither inferences applied in a hypothesis test (significance) nor inferences applied in an interval estimation are valid. A possible misunderstanding, among researchers in medicine, of why we use random allocation is that we do so to achieve empirically balanced groups at baseline. The purpose of random allocation in RCTs is not to achieve empirically balanced groups at baseline.

Instead, the goal is to substantiate the assumption of equal allocation probability. If equal allocation probability cannot be assumed, then the estimator will be biased (equation 1) as to the extent to which allocation probability covaries with potential treatment outcomes.[1]

$$bias(\bar{Y}) = E(Y - \mu) = \sigma_{PY} \tag{1}$$

And when the estimator is biased, the textbook interpretation of a confidence interval, the most common way to quantify uncertainty in medical research, is no longer valid.

Random allocation, for sure, is no guarantee of equal allocation probability, since sources of bias affect the scientific process as a whole (including the reporting). But without random allocation it might be difficult to assume equal allocation probability.

1. The notion of potential outcomes is explained in detail in Imbens and Rubin 2015.

If inferences are confined to the study sample as such, then all confounders are contained within the sample. But suppose that the goal is to infer to properties of a target population. A non-random sample, which is a common procedure in medical research, could mean that equal sampling probability cannot be assumed, and this would affect confounders (known as well as unknown). Possible confounders may therefore affect this sampling probability. Probabilities of this kind, which may lead to bias, have been modelled, for example, by Cole and Stuart (2010). When sampling probability is equal, the mean of a discrete variable in a finite target population can be calculated as a sum containing this probability (equation 2).

$$\sum_{i=1}^{N} P_i \cdot Y_i = \mu = \frac{1}{N} \sum_{i=1}^{N} Y_i \qquad (2)$$

Yet, the sampling probability of the patients may not be equal. For various reasons, the sampling process may be distorted, so that it is higher for those with better treatment outcomes than it is for those with worse outcomes. This problem is discussed by Worrall in a later article (2007). In this case the weighted sum of probabilities will be higher than the mean treatment outcome in the population, the parameter value (equation 3).

$$\sum_{i=1}^{N} P_i \cdot Y_i > \mu \tag{3}$$

This means that the standard estimator for the mean will be biased, and the textbook interpretation will not be valid.

$$\mu < \sum_{i=1}^{N} P_i \cdot Y_i = E(\bar{Y} - \mu) \tag{4}$$

The problem, from a frequentist perspective, is not whether the treatment and control are perfectly balanced regarding all confounders in an empirical sense. It is whether the assumption regarding equal sampling probability is substantiated or not. For all confounders to be controlled, random allocation is seldom sufficient. Instead, random sampling is required in most cases to substantiate equal sampling probability.

It is an interesting question whether randomization can be superfluous, regarding equal sampling and allocation probabilities, in some cases. Let us look at an equation (5) defining bias as a covariance.

$$bias(\bar{Y}) = \sum_{i=1}^{N} (P_i - \bar{P}) \cdot (Y_i - \mu) \tag{5}$$

It is obvious that equal sampling probability implies that the first factor in equation 5 is zero for all patients. Therefore,

the sum of products will also be zero, and the mean outcome estimator will be unbiased. Random sampling is a way to try to achieve equal sampling probability. In principle, equal sampling probability is possible without random sampling, but this is not likely in medical research. Yet, when this variation is very small, the bias will also be very small, and possibly of no importance.[2]

The second factor can also be zero. This occurs when all patients respond to treatment to the same extent – in other words, when the treatment response is homogeneous across all patients. In this case random sampling is not required for the estimator to be unbiased, and the variation can also be so small that the size of the bias is irrelevant. It may be that in certain cases biochemical processes dominate, so that the variation in patient responses is very small. The random component in the design – random allocation or sampling – may then be less important.

If research can show convincingly that the response variation is negligibly small, the product in equation 5 will be close to zero. This is why randomization and random sampling are superfluous also in a frequentist perspective. This could be the case in the example given by Worrall (2007) of a treatment for persistent pulmonary hypertension (PPHS) using extracorporeal membranous oxygenation

2. The estimator may be unbiased even if sampling probabilities and potential outcomes vary when there is no covariation.

(ECMO). The mortality rate was originally 80%, but after the ECMO treatment had been introduced the rate decreased to around 20%. Worrall (p. 455) explains...

> It was already known that the underlying cause of PPHS was immaturity of the lungs, leading to poor oxygenation of the blood, in an otherwise ordinarily developed baby. Those babies that survived were those that were somehow kept alive while their lungs were developing to maturity. ECMO, in effect, takes over the function of the lungs in a simple and relatively non-invasive way. Blood is extracted from one of the baby's veins before it reaches the lungs, is artificially oxygenated at a membrane, reheated to regular blood temperature and re-infused into the baby's carotid artery, thus bypassing the lungs altogether.

The supremacy of internal validity in medical research

The heavy epistemic weight given to RCTs finds expression in the priority given to internal validity over external validity in medical research. Persson and Wallin (2012, p. 1) state that

> Without a doubt the concepts capture two features of research scientists are aware of in their daily practice. Researchers aim to make correct inferences both about that which is actually studied (internal validity), for instance in an experiment, and about what the results 'generalize to' (external validity). Whether or not the language of internal and external validity is used

> in their disciplines, the tension between these two kinds of inference is often experienced. (p. 1)

This alleged supremacy of internal validity is reflected in Cochrane (www.cochrane.org), one of the most influential networks in medical research. In the Cochrane Handbook (Higgins et al. 2017) bias is only assessed as lacking internal validity (p. 8:2). External validity is described in the handbook as "whether the study is asking an appropriate research question" and is "closely connected with the generalizability or applicability of a study's findings". Internal validity is about "whether it [the study] answers its research question 'correctly'... in a manner that is free from bias". Challenging this priority of internal validity, Persson and Wallin hold that "the two types of validity are deeply intertwined" (p. 2).

As Persson and Wallin indicate, validity can be thought of as the correctness of inferences (p. 1). A typical type of inference in medical research is interval estimation. The flaws that may invalidate interval estimation can be divided into two types. The first type are flaws that produce a covariance of allocation probability and potential outcomes. This would mean that internal validity is not a property of the groups in an experiment, such as empirical baseline balance. Instead it means that the interval estimator is biased, involving expectations that differ from sample parameters.

The second type are flaws that lead to a covariance of sampling probability and potential outcomes. As a consequence of this external validity (e.g. empirical representativeness) is not a property of the sample. Instead, the inference is invalidated as a result of a biased interval estimator, so that expectations differ from parameters in the target population.

Where the estimation addresses a target population in clinical practice the two types of flaw can be collapsed into a single type. The flaws will be related to two steps: first sampling and then allocation. Yet, suppose that the ideal RCT is taken to be based on two random samples from the target population (Lachin 1988). One sample is of treatment and the other of the control. Equal sampling probability in both groups will then mean that the interval estimator is unbiased. Internal and external validity can then be collapsed into a single validity regarding interval estimation as an inference. This supports the claim, made by Persson and Wallin, that internal and external validity can be understood as intertwined, with neither being prior to the other.

Concluding remarks

In this note I have tried to show how random allocation and random sampling can be motivated in medical research within the framework of frequentist statistics. The random component is a way to substantiate the assumption of equal sampling probability (random allocation is a special

case of random sampling), and thereby it contributes to the unbiasedness of the estimators. However, the random component in the design is no guarantee of unbiasedness.

Biased estimators invalidate the most common statistical inferences in medical research such as interval estimation. I have also indicated when random sampling may be superfluous. This can occur when the response to treatment is homogeneous among eligible patients. If this can be motivated, the random component in the designs is not necessary.

Some of Worrall's critique of RCTs is based on Bayesian statistics. This is a different topic, not addressed in my note, since my purpose was to discuss the arguments within a frequentist framework.

References

Cole SR, Stuart EA. (2012). Generalizing Evidence from Randomized Clinical Trials to Target Populations. American Journal of Epidemiology. 172(1):107–115.

Higgins J, Altman D, Sterne J. (2017). Assessing risk of bias in included studies. Chapter 8 in Higgins, J., Churchill, R., Chandler, J., Cumpston, M., (eds.). Cochrane Handbook for Systematic Reviews of Interventions version 520 (updated June 2017): Cochrane.

Imbens, G.W., Rubin, D.B. (2015). Causal inference in statistics, social, and biomedical sciences: Cambridge University Press.

Lachin, J.M. (1988). Statistical Properties of Randomization in Clinical Trials. Controlled Clinical Trials. 9:289–311.

Persson, J., Wallin, A. (2012). Why internal validity is not prior to

external validity. Philosophy of Science Association. 23rd Biennial Meeting (San Diego, CA): PSA 2012 Contributed Papers.
Worrall, J. (2002). What Evidence in Evidence-Based Medicine? Philosophy of Science, 69:316–30.
Worrall, J. (2007). Why There's No Cause to Randomize. British Journal of Philosophy of Science, 58:451–488.

Weather experiences and perceptions of climate change

WÄNDI BRUINE DE BRUIN

Climate experts have long warned that global temperatures are on the rise, and that heatwaves will become more common. To estimate how much climate change contributes to an extreme weather event, those experts require complex statistical analyses. However, people who are not experts in climate change may turn to their own experiences to judge the extent to which climate change is a concern.

Psychologists have been studying how people use their own personal experiences with weather events to draw conclusions about climate change. They have found that, when temperatures are higher than usual, people tend to be more concerned about climate change. This is in line with a psychological phenomenon referred to as the 'availability heuristic', which describes the tendency to draw more heavily on recent and extreme experiences when judging the likelihood of events. Indeed, recent and extreme experiences tend to be more vivid and therefore more 'available' from

memory. Hence, recent experiences with heat waves and unseasonably hot weather may fuel public concerns about climate change.

These patterns raise questions regarding public concern about climate change in countries with moderate climates, where temperatures do not tend to get very high. I currently live in Leeds (United Kingdom), where I hold a University Leadership Chair in Behavioural Decision Making. In July 2012, I moved to Leeds from Pittsburgh, Pennsylvania, in the United States. As soon as I arrived, it was obvious that summer temperatures were much more modest in Leeds than in Pittsburgh. The maximum temperature in Pittsburgh had been 27C in July 2012, as compared to only 18C in Leeds. Similarly low summer temperatures may be experienced by residents of Sweden, the Netherlands, Canada, and other countries with moderate climates.

If people who live in temperate climates draw on their personal experiences with hot weather to inform their climate change concerns, they may feel relatively unconcerned about climate change. With my colleagues in Leeds, I have been studying public perceptions of weather and climate change in the UK. In the national surveys that my colleagues and I conducted, we did find that people in the UK look to weather patterns to judge the degree to which they should be concerned about climate change. However, our UK participants' climate change concerns were more strongly driven by their perceptions of lifetime changes in

wet weather events, such as heavy rainfall and flooding, than by their perceptions of lifetime changes in hot weather events, such as hot summers and heatwaves. In fact, they perceived wet weather events to have increased much more over the course of their lives, as compared with hot weather events. They also liked these wet weather events a lot less than the hot weather events.

Although I found UK summers to be remarkably refreshing, many people in the UK would prefer warmer weather. In the UK, it is common to refer to higher temperatures as 'good weather' no matter how uncomfortable it gets. National surveys even show that some people in the UK hope that their weather will become warmer in the future. Moreover, our findings suggested that UK residents' positive feelings about heat undermines their willingness to implement heat protection behaviours when it gets hot.

Heat protection behaviours recommended by public health experts include avoiding sun exposure, especially around midday, drinking plenty of water, and avoiding vigorous exercise. We found that individuals who liked hot weather more were less willing to engage in these behaviours. Such positive feelings about heat are also thought to explain why tourists from Northern European countries seek prolonged sun exposure on Southern European beaches – much beyond the number of hours of sunbathing recommended by public health experts.

Although UK residents may enjoy the prospect of warmer

temperatures, other climate change impacts are also expected for the UK. The UK government's climate change risk assessment report has identified climate change impacts related to hot weather and wet weather. The former may include increased air pollution, more people getting sick with heat-related illness, escalating summer energy demands due to heavier use of air conditioning, cities trapping heat and becoming 'heat islands', and overheated infrastructure. Climate change impacts from wet weather may include flooded homes and flooded infrastructure.

Preparations for those climate change impacts may require individuals to take action (e.g. behaviour change, refurbishing one's home) as well as to support government actions (e.g. changes to cities, infrastructure, and energy systems). In line with our finding that many UK residents feel positive about the prospect of increasing UK temperatures, we found less concern about prioritising preparedness for hot weather events than for wet weather events – even if climate experts thought of both as having high priority.

To inform UK residents' decisions about climate change preparedness, interventions may need to address the risks of specific climate change impacts, the recommended preparedness actions, and the barriers to implementing them. Additionally, we have found that even those who enjoy hot weather have had negative experiences with heat, but positive heat experiences tend to be more salient in

memory. Moreover, reminding people of their 'forgotten' negative experiences with hot weather can help to motivate them to implement heat protection behaviours. We aim to develop and test interventions that effectively build on people's experiences to promote public preparedness for climate change in different countries.

Acknowledgments

I gratefully acknowledge funding from the UK Economic and Social Research Council, the Swedish Foundation for the Humanities and the Social Sciences (Riksbankens Jubileumsfond) Program on Science and Proven Experience, and the Center for Climate and Energy Decision Making (CEDM) through a cooperative agreement between the National Science Foundation and Carnegie Mellon University (SES-0949710). I thank Andrea Taylor for her comments on this paper. My research has greatly benefited from collaboration with Suraje Dessai and Andrea Taylor (University of Leeds, UK), Baruch Fischhoff and Gabrielle Wong-Parodi (Carnegie Mellon University, US), Carmen LeFevre (University College London, UK), Kelly Klima (RAND Corporation, US), and Sari Kovats (London School of Public Health and Tropical Medicine, UK).

Further reading

Bruine de Bruin, W., Lefevre, C.E., Taylor, A.L., Dessai, S., Fischhoff, B., Kovats, S. (2016). Promoting protection against a threat that evokes positive affect: The case of heatwaves in the U.K. Journal of Experimental Psychology: Applied, 22, 261–271.

Klima, K., Lefevre, C.E., Bruine de Bruin, W., Taylor, A.L., Dessai, S. (2017). Weather expectations inform willingness to adapt to climate change. Working paper. University of Leeds: Centre for Decision Research.

Lefevre, C.E., Bruine de Bruin, W., Taylor, A.L., Dessai, S., Kovats, S., Fischhoff, B. (2015). Heat protection behaviors and positive affect about heat during the 2013 heat wave in the United Kingdom. Social Science and Medicine, 128, 282–289.

Lefevre, C.E., Bruine de Bruin, W., Taylor, A.L., Dessai, S. (2017). Climate change concerns come rain or shine – A reciprocal causal relationship of perceived weather changes and climate change concerns. Working paper. University of Leeds: Centre for Decision Research.

Taylor, A.L., Bruine de Bruin, W., Dessai, S. (2014). Climate change beliefs and perceptions of weather-related changes in the United Kingdom. Risk Analysis, 34, 1995–2004.

Taylor, A.L., Dessai, S., Bruine de Bruin, W. (in press). Public priorities and expectations of climate change impacts in the United Kingdom. Journal of Risk Research.

Wong-Parodi, G., Bruine de Bruin, W. (2017). Informing public perceptions about climate change: A 'mental models' approach. Science and Engineering Ethics, 23, 1369–1153.

Evidence-based medicine, clinical guidelines, and the role of patient preferences

JOHAN BRÄNNMARK

Two major developments when it comes to guiding decision-making in medicine are (i) an increased emphasis on the importance of autonomy and shared medical decision-making and (ii) the rise of the ideal of evidence-based medicine. The former has been building since the 1970s and the latter since the 1990s, so these are no temporary fads. But to what extent are these developments in alignment with each other?

On at least one of the early canonical accounts of evidence-based medicine, developed by epidemiologists at McMaster University, it is an ideal of how medicine is practiced that places good medical practice in the intersection between three domains: research evidence, clinical expertise, and patient preferences (Sackett et al. 1997). An updated version of this model included the patient's clinical state, the clinical setting, and clinical circumstances as a fourth component and broadened the patient-oriented

domain to also include patients' actions, including the extent to which patients will actually follow physician recommendations (Haynes et al. 2002). In either version, the McMaster conception of evidence-based medicine is not, at its heart, a conception of how research evidence should be compiled and weighted, but rather a conception of how research evidence should be integrated into clinical decision-making.

While the physician can bring clinical experience and knowledge of relevant research evidence to the process of shared decision-making, the patient is the main authority on his or her preferences. Although it should certainly be recognized that preferences are often formed when actually having to make a decision, and thus partly shaped by the exact nature of those circumstances (Epstein & Peters 2009), the practice of evidence-based medicine should still, on this kind of conception, facilitate shared decision-making. At the end of the day, it would, however, be unrealistic to expect of every physician to keep up to date with the research, even in his or her own specific field of expertise, and this means that for evidence-based medicine to function in actual practice there is a need for intermediates on which physicians can rely. Here clinical guidelines can play a crucial role. And while clinical guidelines are not inherently tied to evidence-basing, they are by now almost invariably at least advertised as being evidence-based (Guyatt et al. 2008). Accordingly, clinical guidelines will often be an

important intermediary through which evidence-basing potentially enters into the clinical situation. But the guidelines also involve a move from summarizing research to making recommendations – a move that cannot be made without relying, not just on evidence, but also on values. To what extent could this circumscribe the influence of individual patient preferences? In order to discuss this question, we will first have to say something about the structure of decision-making in general and medical decision-making in particular.

Two kinds of decision-making

While the focus in much of what is written about medical decision-making tends to lie on the patient-physician encounter, one important fact about contemporary medicine is that it is overwhelmingly practiced in an institutional context. This is not just about the steady decline of physicians in private practice in favor of employment at larger healthcare units, but also about the way in which health-insurance systems function, the role of government regulations, how questions about responsibility are handled by the legal system, how medical research and compilations of meta-analyses are conducted, and how the medical technology and pharmaceutical industries operate. These factors (and others as well) all come in degrees in terms of the extent to which they shape which possibilities are live options in the patient-physician encounter and which are

not. And while the exact shape many of these factors will take might vary from country to country, the overall trend seems to be clear: towards an increasing institutionalization and division of labor which ensures that individual physicians will, when meeting their individual patients, proceed to an increasing degree on the basis of a vast number of decisions that have always already been taken by others.

In any real-life decision we can distinguish between two main phases in the decision-making process: deciding on the menu and deciding from the menu. Deliberation takes time and effort, and so we need to limit the number of options we consider; we need a limited menu to choose from. What characterizes the items that are on the menu is precisely that they are the alternatives to which we give closer thought and ultimately decide between. There are two things to note here. One is that in everyday life menu-setting is largely unconscious. At any moment there are countless options that are in principle open to us, but we tend only to notice a very limited number of them. The other is that we can go back and forth between these two phases: on closer inspection we might find that there is no good alternative on the menu and then we can try to think critically about the menu again, and consider which items could possibly be added to it. In one-person scenarios this movement back and forth is fairly straightforward, but in multi-person scenarios it might very well be the case that there is a division of labor, and certain people do the main job of

deciding on the menu, while others do the main job of deciding from the menu. It might still be possible for the latter to add options to the menu, but the opportunities to do so will tend to be significantly more limited than in one-person scenarios.

The argument here is that the growing institutionalization of medicine has increasingly separated these two phases of decision-making – deciding on the menu and deciding from the menu. The proliferation of clinical guidelines is an example of menu-setting. Good menu-setting (in any context) reduces complexity and correctly identifies the best options available. Arguably, the value of reducing complexity can even, at least up to a point, justify the options on the menu simply being *good enough* rather than absolutely the best (although it should also be recognized that where the bar of being good enough is set will depend on the context). Menu-setting cannot, however, be accomplished in a reasoned way without guidance from certain values. In the institutional medical context two such values or concerns stand out. Foremost is *cost-effectiveness*. For instance, in the UK, the National Institute for Health and Care Excellence (NICE) manual for developing clinical guidelines, while focusing primarily on procedures for reviewing research evidence, strongly emphasizes the importance of analyzing cost-effectiveness (2014, Chapter 7). The other value is what might be called *stakeholder approval*. From a purely ethical perspective, this may seem to be of little direct importance,

but from an institutional perspective having stakeholder approval, which can be a matter of engaging both with representatives of different medical professions and specializations and with various patient groups, is very important, and it will be difficult to achieve stakeholder approval without involving stakeholders in the process of formulating the guidelines.

What this means, however, is that it will be hard to formulate clinical guidelines without at least partly preempting the role that individual patient preferences and circumstances could potentially have played: certain value-based assessments will already have been made. Up to a point, this is quite reasonable, since especially cost-effectiveness is not just an intrapersonal issue for the individual patient, but an interpersonal one: to the extent that health care is cost-effective, we will be able to provide more health care for more patients. But in potentially moving towards what is starting to look like a utilitarian cost-benefit calculus there is also a risk that patient autonomy will suffer, so a balance needs to be struck here.

Minimizing preemptive paternalism

Preemptive paternalism is here understood as the act of imposing judgments about what are to count as good health outcomes in setting the menu of choices that will then form the starting-point for discussions between individual patients and the physician(s) with whom they interact

in making decisions about which treatment options to pursue. The relevant judgments can be imposed in several different ways, but formulating clinical guidelines on the basis of cost-effectiveness assessments is clearly one of them. If we value patient choice, and if we believe that individual patient preferences are highly relevant in determining what will count as a good, or at least acceptable, health outcome in the individual case, we shall have reason to seek to minimize this kind of preemptive paternalism.

We can distinguish between two kinds of cost-effectiveness assessment. To begin with we have what might be called *fine-grained* analyses, where every treatment option can be precisely ranked in terms of a common metric and where the standard candidate in a healthcare context (and the one embraced in the NICE manual, although not as something that should be applied mechanically) is *cost per QUALY*, i.e., the mean cost for the treatment option divided by the mean number of quality-adjusted life years that it will buy us – a figure that can then be compared with other possible treatment options. But it is also possible to use a *coarse-grained* approach instead, where rankings of health outcomes are constructed in terms of broader categories – e.g., whether two treatment options typically have roughly the same types of health outcome (in which case, if one is more expensive, it probably should not be on the menu) or whether one treatment option is clearly superior to another (it has significant effects that for most patients are likely to

count as good health outcomes, while the other has marginal effects that are unlikely to count as good health outcomes for most patients). This kind of analysis will focus primarily on removing ineffective treatment options from the menu, as compared with the best option(s) and should therefore typically result in a bigger menu, as compared with what tends to come out of fine-grained assessment, and, hence, a potentially larger role for the preferences of the individual patient to play.

Two things should be noted here. One is that the application of coarse-grained cost-effectiveness assessments will not completely remove the element of preemptive paternalism in the making of recommendations; rather, the point here is that opting for such assessments should allow us to minimize the extent to which they are preemptively paternalistic. The other point is that it should be recognized that using fine-grained assessments, and presumably relying on a QUALY framework, does not necessitate a strong narrowing down of choice menus; however fine-grained the analysis, we might still just use it for making more coarse-grained decisions. At an institutional level, it does, however, seem likely that a fine-grained analysis will exert a certain gravitational pull on our decision-making processes. And in the case of formulating clinical guidelines, this would then mean a tendency to narrow down the number of choices that are live options for physicians and patients, and hence the role that can be played by the individual patient's prefer-

ences in determining which treatment option that is the most suitable one. If we value the latter, it would accordingly seem reasonable to use mainly coarse-grained cost-effectiveness assessments in developing clinical guidelines.

References

Epstein, R.M., Peters, E. (2009). Beyond Information: Exploring Patients' Preferences, JAMA 2009, vol. 302: 195–197.

Haynes, R.B. et al. (2002). Clinical expertise in the era of evidence-based medicine and patient choice, Evidence Based Medicine, vol. 7: 36–38.

Guyatt, G. et al. (2008). How to use a patient management recommendation, in Guyatt, G. Users' guides to the medical literature: a manual for evidence-based clinical practice, 2nd ed. New York: McGraw-Hill.

National Institute for Health and Care Excellence, Developing NICE guidelines: the manual. https://www.nice.org.uk/process/pmg20/chapter/introduction-and-overview [accessed September 2017].

Sackett, D.L. et al. (1997). Evidence-based medicine: how to practice and teach EBM, Edinburgh: Churchill Livingstone.

Integrating expert judgment and statistical prediction

Synthesizing materials with mechanical syntheses

ALEX DAVIS

Solving problems at the boundaries of human knowledge, such as discovering a new material, synthesizing it with desirable properties, and getting it into the market, requires the synthesis of human intuition, science, and statistical models. The science of this process, of discovering materials and turning them into commercial products, is in its infancy. At each step of the way expert judgment must face the facts, captured and quantified using statistical models. But those facts are backward-looking, characterizing events that have occurred, but providing little information about where to go next. In contrast, experts have hunches about where to go, but are not able to process all the data. Expert judgment and statistical models have complementary characteristics, and a *mechanical synthesis,* or a prediction rule that is constructed from expert judgment and the outputs of

statistical models, can help turn the art of discovery-to-market into a science.

Consider the following three case studies: discovering 1) how to synthesize titanium oxide (TiO_2) crystalline structures at low temperatures, 2) how to 3D print soft materials (silicon elastomers) for use in biomedical applications, and 3) how to identify parts appropriate for metal additive manufacturing (MAM) in aerospace. These bridge the gap from basic science (TiO_2 synthesis), to design (3D printing elastomers), to commercialization (aerospace MAM). In each case, the expert is operating at the boundary of human knowledge – that is, our knowledge of what materials will emerge when microwaving TiO_2 at low temperatures, what characteristics of the 3D printer, or the chemical composition of the silicone elastomer, gives it the right physical properties, and what elements of redesign and recombination might make a part useful for 3D printing. The expert must use data, from the success (or failure) of previous TiO_2 synthesis experiments in terms of their match to the desired crystalline structure, on whether the elastomer can be stretched and compressed in ways that are useful for the application given the different parameters used in 3D printing, and on whether parts produced using MAM actually improve the bottom line of the cost of a jet engine. Those data are best characterized by statistical models, then combined with expert judgment to create a path forward.

What evidence is there that combining expert judgment and statistical prediction models (mechanical synthesis) is a good idea? Jack Sawyer (1966) of the University of Chicago produced the first meta-analysis of studies that included mechanical synthesis. He categorized prediction methods based on how data were collected and combined. There were two types of data collection: expert assessments (e.g. interviews, intuitive assessment) and mechanical measures (e.g. response to a survey, trait measure). There were also two types of combination: expert judgment (e.g. prediction on a 0-100 scale) and mechanical combination (e.g. regression). The syntheses (either expert or mechanical) also included the output of other prediction methods and were combined using either expert or mechanical combination. This yielded eight possible configurations, of which mechanical synthesis was compared with five:

1. *Pure expert*: expert data collection, expert combination
2. *Trait ratings*: expert data collection, mechanical combination
3. *Pure statistical*: mechanical data collection, mechanical combination
4. *Expert composite*: mechanical & expert data collection, expert combination
5. *Mechanical composite*: mechanical & expert data collection, mechanical combination
6. *Mechanical synthesis*: mechanical composite plus pure expert and/or expert composite, mechanical combination

In total, 10 studies had mechanical syntheses, where expert judgment and other data were combined mechanically to yield a prediction. Table 1 has a description of the studies. Overall, in 10 studies with 20 comparisons, mechanical synthesis was never inferior to another method, and was superior in 50% of comparisons (and equally good in the other 50%). In 9 of the 10 comparisons available, mechanical synthesis performed better than expert approaches (pure expert = 2, expert composite = 8), and tied statistical approaches in every case (pure statistical = 7, mechanical composite = 2). The result suggests two things: 1) that statistical combination approaches are never inferior, and almost always superior, to expert combination, and 2) that adding expert judgment to a statistical model contributes little, if any, benefit.

One curious finding is that *expert* synthesis, where the expert takes all the data available, including the output of a statistical model, and combines the information to produce a prediction, was no better, but also *no worse*, than mechanical prediction methods. This was the case for Watley and Vance (1964) as well as Melton (1952) in predicting undergraduate grades, as well as Harris (1963) on the winner of football games. One explanation of this result is *model mimicry*, where an expert, when given a mechanical prediction, simply reports the prediction back. This induces perfect dependence between the model and the expert judgment, which can reduce overall system (expert and

Table 1

Authors	Year	Prediction context	Results
Bobbit & Newman	1944	Final class standing in coast guard officer candidate training	mechanical synthesis = pure statistical = clinical composite
Borden	1928	Parole success in NJ	mechanical synthesis > clinical composite
Burgess	1928	Parole violation in IL	mechanical synthesis > clinical composite
Cliff	1958	Naval officer candidate grades	mechanical synthesis = pure statistical > trait ratings = clinical composite
Doleys & Renzaglia	1963	College freshman grades	mechanical synthesis = mechanical composite = clinical composite
Dunlap & Wantman	1944	Pilot trainee potential	mechanical synthesis = mechanical composite = pure statistical > pure clinical
Hamlin	1934	Inmate institution adjustment	mechanical synthesis > clinical composite
Peirson	1958	College freshman grades	mechanical synthesis = pure statistical > pure clinical
Sarbin	1943	College first quarter grades	mechanical synthesis = pure statistical
Westoff	1958	Number of live births for engaged couples	mechanical synthesis = pure statistical > clinical composite

model) performance if experts have knowledge that does not overlap with the statistical model. For example, in the medical domain, Knaus *et al*. (1995) found that expert physicians and a statistical model had *independent* predictive power, and combining their predictions in a mechanical synthesis was better than either alone. This raises the question of how to use both statistical and expert predictions in sequential prediction environments, where new experiments (or patients) must constantly be designed or assessed in sequence. Should the expert be informed about the statistical model if a knowledge of the output of the model induces mimicry, potentially reducing overall performance? Can the expert infer the output from the actions suggested by a statistical model? Existing studies are not yet able to answer these questions.

Another issue is that studies of mechanical synthesis have followed a single process: experts make judgments, other data are collected, statistical models are constructed, then a mechanical synthesis combines them to produce a final prediction. In situations where there are relatively few factors to consider – as in the case of TiO_2 synthesis, where the microwave has only a few parameters – the mechanical composite approach makes sense: let experts make predictions and combine that with models of prior data. But this is not the only type of problem, nor the only reasonable approach. In 3D printing of silicone elastomers, for example, the space of possible parameters is huge. To address this,

the expert and statistical model can alternate, with the expert choosing the space of parameters to search (e.g. chemical concentrations, speed of the extruder) and the model choosing experiments within that space based on prior data and effective search algorithms.

Computational power, statistical models, and artificial intelligence, have become ubiquitous. Mechanical systems can solve many narrow problems, from calculating *pi* to many digits, to predicting what advertisement people are most likely to click. Harder problems, where there are few data on which to build models and my factors (possibly undefined) determine outcomes, are currently intractable for mechanical approaches. Experts rely on their knowledge and judgment to address these problems, but they may miss patterns that mechanical models could pick up (with the expert's help), or not trust their own judgment if they think artificial intelligence is better than their own. Systems that use human experts and machines optimally do not exist, and represent another boundary of human knowledge. Crossing that boundary will require us to find ways of using machines without undermining experts, and to rely on experts without ignoring the data.

References

Burgess, E.W. (1928). Factors determining success or failure on parole. In A. A. Bruce (Ed.), The workings of the indeterminate

sentence law and the parole system in Illinois. Springfield: State of Illinois, pp. 203-249.

Cliff, R. (1958). Validation of selection procedures in enlisted-to-officer programs. USN Bureau of Naval Personnel Technical Bulletin, No. 58-11.

Doleys, E.J. Renzaglia, G.A. (1963). Accuracy of student prediction of college grades. Personnel Guidance Journal, 41, 528–530.

Dunlap, J.W., Wantman, M.J. (1944). An investigation of the interview as a technique for selecting aircraft pilots. Washington, D. C.: Civil Aeronautics Administration. (Report No. 33)

Hamlin, R. (1934). Predictability of institutional adjustment of reformatory inmates. Journal of Juvenile Research, 18, 179–184.

Harris, J.G. (1963). Judgmental versus mathematical prediction: An investigation by analogy of the clinical versus statistical controversy. Behavioral Science, 8, 324–335.

Knaus, W. A., Harrell, F. E., Lynn, J., Goldman, L., Phillips, R. S., Connors, A. F., et al. (1995). The SUPPORT prognostic model: Objective estimates of survival for seriously ill hospitalized adults. Annals of Internal Medicine, 122(3), 191–203.

Melton, R. S. (1952). A comparison of clinical and actuarial methods of prediction with an assessment of the relative accuracy of different clinicians. Unpublished doctoral dissertation, University of Minnesota.

Pierson, L.R. (1958). High school teacher prediction of college success. Personnel Guidance Journal, 37, 142–145.

Sarbin, T. R. (1943). A contribution to the study of actuarial and individual methods of prediction. American Journal of Sociology, 48, 593–602.

Sawyer, J. (1966). Measurement and prediction, clinical and statistical. Psychological Bulletin, 66(3), 178–200.

Watley, D.J, Vance, F.L. (1964). Clinical versus actuarial prediction of

college achievement and leadership activity. United States Office of Education Cooperative Research Project No. 2202, Sept. 1964, University of Minnesota.

Westoff, C.F., Sagi, P.C., Kelly, E.L. (1958). Fertility through twenty years of marriage: A study in predictive possibilities. American Sociological Review, 23, 549–556.

VBE and the PhD

BARRY DEWITT

I first visited Lund as a third-year doctoral student from Carnegie Mellon University. I recall asking a group of philosophers about the difference between "applied philosophy" and "theoretical philosophy." The two terms were new to me when I began my interaction with the VBE group, and my hosts in Sweden explained to me (with a touch of humor) that applied philosophers theorized about applied problems, and theoretical philosophers theorized about theoretical problems.[1]

I was, of course, surrounded by both applied and theoretical philosophers, including Johannes Persson. I had arrived in Lund as a student of behavioral decision research, a discipline combining psychology, mathematics, statistics, and elements of many other subjects. I had previously been

1. The explanation also involved a story about the origin of the distinction between the two types of philosophy that centered on a clever strategy to receive more funding for philosophy departments – I do not know if the story was told at my expense, but I have repeated it many times since, to equally naïve North American non-philosophers.

a student of pure (i.e. *theoretical*) mathematics, and had spent my time worrying about things like Springer varieties and von Neumann algebras (Dewitt & Harada, 2012; Döring & Dewitt, 2014). My decision to change disciplines was based on a desire to worry about things whose ontological status was clearer (I hoped).

On being accepted into the Department of Engineering and Public Policy's PhD program at Carnegie Mellon to work with Baruch Fischhoff, I had several months before my studies officially began. To ease my transition, Professor Fischhoff provided me with a reading list, which included two works that set the stage for the subsequent four years (and counting).

The first was Coombs, Dawes, and Tversky's *Mathematical Psychology: An Elementary Introduction* (Coombs, Dawes, & Tversky, 1970). Before I met Professor Fischhoff, others had told me that if I wanted to use my mathematical education and apply it to something that concerned policy problems, the best options were economics and statistics. Coombs et al. introduced me to an entire discipline I did not know existed, one that intersected with both economics and statistics and included much more. That book led me to learn more about Tversky's work in mathematical psychology via the *Foundations of Measurement* series (Krantz, Luce, Suppes, & Tversky, 1971). At the time, I was surprised to discover theorems and proofs concerning measurement that looked exactly like some of those I had

seen in mathematical physics. I did not know that Patrick Suppes, for example, had done fundamental work not only in quantum theory, but also in decision theory! The fact that Tversky co-wrote the *Foundations* and also co-invented prospect theory made me hopeful that I had in fact found the balance between the theoretical and the applied that I was seeking.

Professor Fischhoff had also suggested *Decision, Probability, and Utility: Selected Readings*, a volume edited by Peter Gärdenfors and Nils-Eric Sahlin (1988). It was my first dive into decision theory. One topic that stands out four-and-a-half years later is the presentation and discussion of Savage's sure-thing principle. I wondered why so many would spend so long debating it. I learned of the importance of the normative-descriptive-prescriptive framework while working through some of the arguments for and against the principle: What should a decision theory be? How does one describe actual decisions? How does one bridge the two, when they do not match? Finding another intellectual outlet that appreciated and used mathematics was gratifying, but it was the "structural" feel of these three questions that most reminded me of the aesthetic of mathematics.

Once I began the PhD program, I learned more decision theory and mathematical psychology, chiefly under the tutelage of Professors Baruch Fischhoff, Alexander Davis and Stephen Broomell. I also came to understand much

more about how one might do "real-world" applied research. Our first work took some of the prototypical mathematical psychology models – signal detection theory and multidimensional scaling – and applied them to the risk perception of natural hazards (Dewitt, Fischhoff, Davis, & Broomell, 2015).

While studying pure math, I had dismissed applied projects. In trying to apply theory to data, I learned how naïve I had been – good applied research is just as rewarding as theoretical work, and also involves the latter if one is lucky. The applications of the VBE group exemplify this approach: studying science and proven experience in action provides an opportunity to improve decision-making (e.g. of physicians in clinical encounters) while learning more about how an important expert group diverges from some normative decision theory (e.g. expected utility theory).

My interaction with the VBE group began after my first work on natural hazards and coincided with a graduate seminar in the philosophy department at Carnegie Mellon, led by Professor Teddy Seidenfeld, on Savage's *Foundations of Statistics* (Savage, 1972). That semester, which also included meeting Professor Janel Hanmer at the University of Pittsburgh Medical Center, changed the course of my doctoral education. I began working on applications in healthcare, where expert judgment and science often need to be synthesized to create informed policy.

My first experience of that synthesis occurred during the

seminar on Savage. My term paper examined the foundations of the measurement of preference-based health-related quality of life (Dewitt, Davis, Fischhoff, & Hanmer, 2017), a metric that is meant to describe the utility of health and is used as an input to many health policy decisions. The formal foundation built by decision theory (e.g. from philosophers including Rawls and Sen) provided a lens with which to examine the choices of applied researchers, who were constructing the measurement tools used in policy analysis. We argued that current conventions ignored many of the insights of the philosophers, and we suggested a procedure for implementing those insights that was informed by interventions based on behavioral decision research and its normative-descriptive-prescriptive framework (Fischhoff, 2015; Stern & Fineberg, 1996).

However, in the process of translating philosophical and decision-theoretic concepts to an applied context – health policy analysis – we were required to make compromises that would be controversial to philosophers. When I went from my engineering department to the philosophy department to seek the input of professional philosophers, they remarked on the non-trivial ethical problems raised by our proposal. I like to think that in revealing the theoretical flaws of an applied tool – and suggesting how one might shore them – we achieved some measure of "applied-philosophical" research success, even if our suggestions raise ethical questions with no clear answer. In some purely

theoretical pursuits the posing of clear research questions has proven as important for the field as their solutions (e.g. Hilbert's problems). In contrast, I have learned that people in the applied domains are often willing (and able) to engage in theoretical discussion if its practical implications are clear and they are offered a way to reach a workable, if imperfect solution.

My involvement with the VBE group has been a formative part of my graduate education. At our last meeting, in Pittsburgh, when Professor Sahlin announced this volume in honor of Johannes Persson, he remarked with incredulity on the number of decades that had passed since their first meeting. It occurred to me that their (often joint) research has swung from the theoretical to the applied, often containing elements of both. I hope that my career might be similarly diverse. I am grateful to the VBE project for setting me on a path where that outcome has high probability.

References

Coombs, C. H., Dawes, R. M., Tversky, A. (1970). Mathematical Psychology: An Elementary Introduction. Prentice-Hall.

Dewitt, B., Davis, A., Fischhoff, B., Hanmer, J. (2017). An Approach to Reconciling Competing Ethical Principles in Aggregating Heterogeneous Health Preferences. Medical Decision Making, 0272989X1769699. http://doi.org/10.1177/0272989X17696999

Dewitt, B., Fischhoff, B., Davis, A., Broomell, S. B. (2015). Environmental risk perception from visual cues: the psychophysics of

tornado risk perception. Environmental Research Letters, 10(12), 124009. http://doi.org/10.1088/1748-9326/10/12/124009

Dewitt, B., Harada, M. (2012). Poset pinball, highest forms, and (n-2, 2) Springer varieties. Electronic Journal of Combinatorics, 19, 1–35.

Döring, A., Dewitt, B. (2014). Self-adjoint Operators as Functions I: Lattices, Galois Connections and Spectral Order. Communications in Mathematical Physics, 328, 499–525. http://doi.org/10.1007/s00220-014-1991-3

Fischhoff, B. (2015). The realities of risk-cost-benefit analysis. Science, 350(6260), aaa6516-aaa6516. http://doi.org/10.1126/science. aaa6516

Gärdenfors, P., Sahlin, N.-E. (Eds.). (1988). Decision, Probability, Utility: Selected Readings. Cambridge University Press.

Krantz, D., Luce, R. D., Suppes, P., Tversky, A. (1971). Foundations of measurement, Vol. I: Additive and polynomial representations. Academic Press.

Savage, L. J. (1972). The Foundations of Statistics (Second Revised Edition). New York: Dover Publications, Inc.

Stern, P. C., Fineberg, H. V. (Eds.). (1996). Understanding Risk: Informing Decisions in a Democratic Society. National Academy Press.

The art of proven experience

BARUCH FISCHHOFF

The sciences analyze options, relying on the tools of their disciplines. The arts create options, relying on the proven experience of their practitioners. When successful, decision science bridges the sciences and the arts, treating the two worlds as equal, essential partners to sound decision making.

These two worlds are largely mysteries to one another. Herbert Simon recognized that mutual opacity when he used design processes to illustrate the limits to rational decision making. As he observed, designers can create an unlimited set of options with wildly diverse outcomes whose realization can depend on myriad intervening events. The resulting decision space (defined by those options, outcomes, and events) can be so complex as to defy exhaustive evaluation.

Simon's two proposed strategies for managing such complexity rely on heuristics. Although the heuristics that interested him (e.g. ones for locating warehouses or playing chess) were not grounded in philosophical inquiry, they had

properties of ones that are: broadly applicable, holistic rules, formulated in sufficiently precise terms that their implications and merits can be productively evaluated.

Thus, heuristics provide a bridge that allows applying products of the sciences to evaluate products of the arts. Those who generate heuristics have neither scientists' commitment to generality nor artists' commitment to idiosyncrasy. Rather, they are committed to the wisdom of finding reasoned, imperfect solutions to complex problems. Nowhere is that clearer than in philosophers' deliberations over the merits of alternative principles, probing their applicability, often to imaginatively crafted situations.

Heuristic strategies

One of Simon's two strategies, satisficing, employs heuristics to search for potentially acceptable options, which are then evaluated in terms of their expected performance on all valued outcomes. It relies on substantive knowledge regarding where good options might be found. Simon's own work, collaborating with Allen Newell and others, embraced this strategy. It adopted the simplifying research heuristic of examining decisions (e.g. chess moves) where the ultimate evaluation criterion was clear, even if the value of intermediate decisions was not.

Simon's second strategy, bounded rationality, employs heuristics that ignore enough options, outcomes, or uncertainties, to render a decision problem compact enough to

allow thorough evaluation. Heuristics for ignoring options and uncertainties rely on substantive knowledge regarding which options might work and which uncertainties might affect a decision's impacts. Heuristics for ignoring outcomes rely on ethical concerns regarding which issues cannot be compromised. (As a practical matter, one might apply substantive knowledge to screen for outcomes that do not vary enough across the remaining options to affect the choice. However, the act of checking reflects an ethical commitment to seeing whether an ignored outcome might matter.)

Decision science has no special expertise in identifying the outcomes that should matter. Indeed, its expression in neoclassical economics is grounded in indifference to matters of taste, holding that what people value is their own business. Revealed preference analyses attempt to discern what those values are – employing the simplifying research heuristic that decision makers pursue them rationally. Additional assumptions (e.g. about efficient markets and symmetrical information sharing) allow economists to edge toward treating those preferences as appropriate as well.

In contrast, decision science as practiced by psychologists assumes that people are imperfect. They might neglect attractive options, misjudge uncertainties, and not know what they want – perhaps being confused, perhaps being mistaken, perhaps being misled. In order to cope with that complexity, psychologists employ the simplifying research

heuristic of examining behavior in experimental settings, offering choices with limited options and clear contingencies, all carefully explained.

Heuristic insights

The heuristic strategies of these disciplines complement one another in ways that could make their current edgy rapprochement mutually beneficial. The satisficing strategy that guides the top-down approach of revealed preference economics might identify broadly observable regularities (e.g. discounting future outcomes). The bounded rationality strategy that guides the bottom-up approach of experimental psychology might study the dynamics of those regularities (e.g. whether people care less about future outcomes or are uncertain about receiving them).

However, neither discipline has any inherent insight into the primitives of those decisions. What are the creative processes determining which options come into being? How well can a future be known? What is really at stake in a choice? What passions are, might, or should be evoked? These issues require philosophy, applying its analytical tools to examine the heuristics that guide our disciplines when examining decisions about objects that the artists of the world, broadly defined, create.

Doing so might be a modest departure from the normal work of philosophers, who are accustomed to creating unnatural situations so as to understand proposals applied

to them. As a result, it might prove a useful enterprise. Whether it does will depend on how patient philosophers are with scientists' inertia, able to domesticate new notions at a slow pace, lest they forfeit the proven experience of their craft. In this light, philosophers' role is that of bilateral problem feeders,[1] suggesting directions for scientists, learning something from what scientists can and cannot absorb.

One problem loop

In 1969, a nuclear physicist and technologist, Chauncey Starr,[2] presented a set of highly aggregated estimates of the risks and benefits to society from several activities and technologies, including hunting, railroads, natural disasters, and the Vietnam War. Assuming a rational society, which got what it wanted in terms of risk-benefit tradeoffs, Starr argued that his analyses showed that that society had accepted risks that rose with the cube of the benefits and that were three orders of magnitude larger for involuntarily incurred risks (e.g. commercial aviation) than they were for voluntarily incurred ones (e.g. general aviation) for any given level of benefit.

Feeding off a problem in that formulation, a philosopher,

1. Thorén, H., Persson, J. (2013). The philosophy of interdisciplinarity: Sustainability and problem-feeding. Journal of General Philosophy of Science, 44, 337–355.

2. Starr, C. (1969). Social benefit versus technological risk. Science, 165, 1232–8.

William Lowrance,[3] dissected the different ethical principles that might motivate such a double standard (were it actually achieved). He also noted other qualitative features (e.g. whether risks are well understood, evoke a feeling of dread, take lives catastrophically) that might also prompt a double standard, and presented nuanced dissections of their ethical standing.

Feeding off a problem in that formulation, my colleagues and I[4] asked whether people actually did see societal risk-benefit tradeoffs as reflecting their preferences (no), whether they were willing to accept more risks in return for greater benefits (yes), whether they wanted more benefits from involuntary risks (yes), and whether they wanted double standards for other qualitative features of risk (yes). As a further complication, these empirical results also showed that the qualitative features are correlated. For example, involuntary risks tend to be new and unknown to science. As a result, even a clear double standard for risks with these features might not have a clear source without independent evidence regarding which feature drove it.[5]

3. Lowrance, W.W. (1976). Of acceptable risk: Science and the determination of safety. William Kaufman, Los Altos, CA.

4. Fischhoff, B., Slovic, P., Lichtenstein, S., Read, S., Combs, B. (1978). How safe is safe enough? A psychometric study of attitudes towards technological risks and benefits. Policy Sciences, 9, 127–152.

5. Fox-Glassman, K., Weber, E.U. (2016). What makes risk acceptable? Revisiting the 1978 psychological dimensions of perceptions of technological risks. Journal of Mathematical Psychology, 75, 157–169.

Feeding forward, the artist Adele Henderson[6] devoted her "Recent Works" to riffs on these results, imagining the meaning of the objects, the features, and the physicist's vision that fed problems to philosophers and psychologists. Her vision awaits the philosopher who will see in it the problem that can continue the feeding process, asking psychologists to apply their science and proven experience to heuristic observations extracted from Henderson's art.

6. http://www.adelehenderson.com/1recentwork/index.html

Dimensions of science and proven experience and variants of evidence-based medicine in practice

CHARLOTTA LEVAY

Introduction

The Swedish legal precept that healthcare should be conducted in line with 'science and proven experience' may be seen as a local, historically rooted counterpart to evidence-based medicine, the currently dominant approach to medical practice which emphasises that decisions about the care of individual patients should be based on the best available evidence from clinical research. There are clearly parallels between the two frameworks. The 'science' component of 'science and proven experience' is often interpreted in accordance with the principles set out by evidence-based medicine, privileging randomised controlled trials over other types of evidence. As for 'proven experience', proponents of evidence-based medicine explicitly recognise that in clinical care, research evidence needs to be integrated

with individual expertise acquired through clinical experience (e.g. Sackett 1997).

Still, as shown by Johannes Persson and colleagues (forthcoming), there are noteworthy differences between the two sets of concepts, especially when it comes to the relative importance given to experience. In the evidence-based medicine model the clinician's experience is primarily a means of applying research evidence to particular cases, but in the science and proven experience model the experience that the clinician adds to decisions can also qualify as relevant evidence. Moreover, the experience recognised in the evidence-based model is explicitly that of the individual clinician, whereas the 'proven experience' implied in the science and proven experience model has a collective dimension and may also refer to knowledge that is accepted within a group of practitioners (Wahlberg and Persson 2017).

In this chapter, I will attempt to describe a new way of relating the two conceptual frameworks to one another. Drawing on recent case studies of practitioners' work in everyday healthcare settings, I will consider different ways in which evidence-based medicine can play out in practice. My suggestion is that the main variants can be fruitfully characterised in terms of distinctive ways of using science and distinctive types of proven experience deployed in the process. That is to say, I will show how the framework of science and proven experience may help qualify and con-

trast significant variants of evidence-based medicine in practice.

Dimensions of applying science: Following guidelines or critically assessing evidence

In an in-depth interview study of paediatric residents at two US academic hospitals all of the residents reported that, at least occasionally, they 'did' evidence-based medicine (Timmermans and Angell 2001; Timmermans and Berg 2003: 142–165). Their practice with patients was guided by experienced attending physicians who made final clinical decisions, but the residents were regularly encouraged to consult the literature, and they indicated that when encountering a new situation or dilemma they could prepare by looking at guidelines or primary research before talking with the physician. Remarkably, the residents took sharply different approaches to locating and using evidence. They displayed two key orientations to evidence-based medicine: most relied on available literature as *librarians*, while a few of them evaluated it critically as *researchers*.

For librarian residents evidence-based medicine involved pragmatically relying on authoritative literature to quickly solve a diagnostic or treatment problem at hand. They chiefly sought out pre-packaged evidence in secondary sources such as handbooks, practice guidelines and review articles. When librarian residents read review articles, they skimmed the methodology and focused on the conclusion

and research findings. They found much evidence via MD Consult, a user-friendly and readily available database, thus avoiding the five-minute walk to the library.

For researcher residents, by contrast, evidence-based medicine entailed active evaluation and interpretation of the literature. They looked in the literature not just for pragmatic guidance but for a variety of factors to take into consideration during decision-making. Wary of guidelines that merely expressed the consensus of experts in the field, researcher residents sought out findings from randomised controlled trials and meta-analyses with critical assessments of the available evidence. They used guidelines and review articles only as intermediary steps to more specialised evidence and consulted a variety of databases to get a more complete overview of the topic, even if that required an extra trip to the library.

Interestingly, these findings correspond with the two main variants of evidence-based medicine in circulation. In common medical parlance, evidence-based medicine primarily means the use of clinical practice guidelines to disseminate proven diagnostic and therapeutic knowledge (Timmermans and Berg, 2003: 3). This implies that the primary role of individual clinicians is to consult and follow guidelines that are ideally evidence-based, just as the resident librarians did. However, the original idea of the founders of evidence-based medicine was that each clinician should seek out and critically evaluate evidence

him- or herself (Daly 2005: 90–91), as the resident researchers did.

Proponents of evidence-based medicine eventually accepted that most clinicians will use secondary sources, since they lack the time and interest to acquire the skills needed to synthesise primary literature. This acknowledgement was set out in a 2000 editorial. Interviewed later about the change of view the evidence-based medicine pioneer Gordon Guyatt explained:

> When I started, I thought we were going to turn people into *evidence-based practitioners*, that they were really going to understand the methodology, that they were really going to critique the literature and apply the results to clinical practice. I no longer believe that. What I believe now is that there will be a minority of people who will be evidence-based practitioners, and that the other folk will be *evidence users* who will gain a respect for the evidence and where it comes from and a readiness to identify evidence-based sources which summarize the evidence for them. But they are not actually expected to read and understand the articles and really be able to dissect the methodology.
>
> Gordon Guyatt (cited in Daly, 2005: 91, emphases added).

In practice, then, there appear to be two main ways for clinicians to ensure that their decisions are based on the best available evidence, and they can be conceptualised as two different ways of drawing on science in clinical care: the

clinician can either follow guidelines, acting as a 'librarian' or evidence user, or critically assess the evidence, acting as a 'researcher' or evidence-based practitioner – see the left-hand boxes in Figure 1.

		PROVEN EXPERIENCE	
		Personal	Collective
SCIENCE	Following guidelines	• 'Librarian' clinicians • Evidence users	• Collegial meetings for following guidelines
	Critically assessing evidence	• 'Researcher' clinicians • Evidence-based practitioners	• Collegial meetings for assessing evidence

Figure 1. Dimensions of science and proven experience and variants of evidence-based medicine in practice.

Dimensions of proven experience: Personal or collective

The evidence-based medicine model recognises that clinicians need to employ their own clinical experience when applying scientific evidence in patient care, as already mentioned. But physicians also discuss treatments with each other, which means that they draw on collective experience. In hospital care, decision-making is largely a protracted, collectively organised activity marked by continual debates, negotiations, and revisions, with reference to precedents as well as research findings (Atkinson, 1995: 52–58). The pro-

cesses involved are dispersed in time and space and distributed across several teams and individuals, and a good deal of work gets done in the course of collegial talk.

In cancer care, collegial discussions have a central place. Clinicians of different disciplines need to coordinate their efforts, and multidisciplinary conferences have become a standard component of clinical decision-making. In a study of multidisciplinary meetings in French cancer care, Castel (2008) notes that this constitutes a collective approach to medical care. Given the rapid development of their field, most cancer physicians saw it as essential to rely on up-to-date, evidence-based guidelines. Multidisciplinary meetings were occasions for lateral control among peers, where doctors reminded each other of existing recommendations and new evidence. They used the meetings to ascertain that others and they themselves had considered all relevant factors in complex decisions. Sometimes they saw a need to depart from guidelines, especially when treating older patients, and by discussing such departures in multidisciplinary meetings they could either make sure that they had the support of other physicians or change their minds and revert to standard treatment. In one in four of the 200 decisions investigated in the study, the final decision differed from what the presenting doctor had proposed. Doctors with extensive experience and those who could motivate their decisions with reference to relevant research literature enjoyed particular confidence during multidiscipli-

nary conferences, while doctors with leading positions and university credentials could have their proposed decisions questioned and reversed.

The study of French multidisciplinary cancer meetings fleshes out what a collective dimension of 'proven experience' might consist of, beyond mere reliance on generally established practice – a reliance harshly criticised by the advocates of evidence-based medicine. The collective approach represents a systematic way to convene and draw on the experience of several clinicians with relevant, complementary expertise to ensure that guidelines are followed judiciously. This variant of evidence-based medicine in practice can be characterised as applying science by following guidelines with the help of collective proven experience – see the upper-right box in Figure 1.

Finally, collective experience can also be mobilised in joint critical assessment of evidence. Discussion in the French multidisciplinary meetings investigated by Castel (2008) included references to new research studies, and, in a more indirect way, the meetings stimulated doctors to immerse themselves in the literature in order to appear competent to other participants. In addition, other studies of multidisciplinary cancer conferences (Frykholm and Groth, 2011) and of collegial talk more generally (Atkinson, 1995: 58) have reported that doctors exchanged information about the latest research findings and debated how to interpret the available evidence. This final variant of evidence-

based medicine constitutes another, distinctive combination of science and proven experience, implying that science is drawn on by critically assessing evidence through collegial discussions – see the lower-right box in Figure 1.

The empirical studies discussed here illustrate that evidence-based medicine in practice is not the monolith it is often assumed to be. The ease with which the main variants can be described in terms of science and proven experience suggests that further exploration of this older, Swedish concept can shed new light on a major trend in current international medicine.

References

Atkinson, P. (1995). Medical Talk and Medical Work: The Liturgy of the Clinic. London: Sage.

Castel, P. (2008). La gestion de l'incertitude médicale: approche collective et contrôle latéral en cancérologie. Sciences Sociales et Santé 26(1): 9–32.

Daly, J. (2005). Evidence-Based Medicine and the Search for a Science of Clinical Care. Berkeley: University of California Press.

Frykholm, O., Groth, K. (2011). References to personal experiences and scientific evidence during medical multi-disciplinary team meetings. Behaviour and Information Technology, 30(4): 455–466.

Persson, J., Vareman, N., Wallin, A., Wahlberg, L., Sahlin, N.-E. (forthcoming). Science and proven experience: a Swedish variety of evidence-based medicine and a way to better risk analysis? Journal of Risk Research.

Sackett, D. L. (1997). Evidence-based medicine. Seminars in Perinatology 21(1): 3–5.

Timmermans, S., Angell, A. (2001). Evidence-based medicine, clinical uncertainty, and learning to doctor. Journal of Health and Social Behavior 42(4): 342–359.

Timmermans, S., Berg, M. (2003). The Gold Standard: The Challenge of Evidence-Based Medicine and Standardization in Health Care. Philadelphia: Temple University Press.

Wahlberg, L., Persson, J. (2017). Importing notions in health law: science and proven experience. European Journal of Health Law 24: 1–26.

It's values that matter

NILS-ERIC SAHLIN

Baruch Fischhoff, one of the members of the VBE research programme, recently had a close encounter with a self-driving Swedish car. It took place in the US at an intersection in Pittsburgh. The autonomous car, blocked by a road crew and a concrete mixer, had to make a right turn. Fischhoff, did nothing, waiting on the sidewalk for the car to move. Fischhoff's decision not to move was autonomous. It was a decision based on knowledge, information and preferences: it was he who should give way to the car, not the other way round. An informed guess would be that Fischhoff also had a desire not to be knocked down by the car. The car too did nothing. We have no idea how the car's computer assessed the situation. Maybe it was programmed to stop and stand still in the circumstances that had arisen. The deadlock was broken by a human being in the car – an example of what sometimes is called "meaningful human control".

It is predicted that "AI and machine learning, robotics and mechanical engineering, mechatronics, systems with various degrees of autonomy will become available on a

large scale in the coming years" (Santoni de Sio & van den Hoven). The aim is to create more or less autonomous artificial agents or systems which, without human intervention or aid, can perform various tasks. We are talking about self-driving cars, robotic nurses, automated surgeon-free surgery, companion robots... and autonomous armed drones.

Already, surgeons use computer-controlled robots to assist them in some surgical procedures. It is said that robotic surgery leads to reduced pain, faster recovery times, and a lowered risk of blood loss and infection. However, studies have shown that surgeons do as well as the robots we have today. The goal, of course, is to develop robots that perform far better than they do currently – to design robotic surgeons that are a hundred per cent autonomous. But then robotic surgeons, and autonomous cars and drones, have to take their own decisions. Decisions, in other words, that are not controlled by human beings – that is what makes the "bots" independent and self-governing. What then is a decision?

Decisions involve a combination of knowledge/information and preferences/values. An autonomous agent, a robot, cannot take a decision unless the knowledge and information it has is quantified, for example in terms of probabilities. Likewise the preferences and values have to be dressed up in numbers, for example in terms of utilities. But probabilities and utilities, however precise and robust, make

no decision. They are inert. What is needed is a rule combining the quantified uncertainties and values. A problem here is that there are many decision rules. Which one to choose is an open question. And among other things, the choice depends on how the knowledge and values we have are quantified.

Of course, this is not how we, you and me, take everyday decisions, if any. It is a sketch of how more or less ideal agents, well-bred robots, ought to take decisions in specific situations.

Do self-driving cars take decisions? Do non-human robotic surgeons take them? Who is to be blamed if an autonomous car takes the "wrong" decision and kills someone? Who is accountable if a robotic surgeon kills a patient, or when a lone nursing bot seriously harms an individual in the recovery ward? My aim in this short note is to raise a couple of questions that I think need to be addressed. In one way or another, the questions all have to do with science and proven experience.

Information is not enough

Information on its own makes no decision. Does a self-driving car on the streets of Pittsburgh with no preferences or values take decisions? No, definitely not. What kind of information controls the car's movements? The car, let's say, has detailed maps of the city, and other information, information gathered in feedback loops when it and other cars have

driven, or have been driven, around Pittsburgh. The car is programmed to use the information to make inferences. Inferences, however, are not decisions. Without preferences and values, the car wouldn't move a metre. The question is – whose preferences, and whose values, make the car move?

Algorithms tell the car to drive from one place to another, to stop when an obstacle blocks its way. Someone has decided that the information the car is relying on is the information that is important, and that other things can be ignored. In itself, that means that preferences and values have already been added. Someone has made sure that the car does not bump into a human-like object. Why? Because we like that type of object more than we like, for example, lamp posts and trees. But who are "we"? The scope of the quantifier is not to be neglected.

Big Data and Deep Learning are two areas attracting scientific attention. Many organisations have been collecting gargantuan quantities of information. Depending on which domain we are talking about, these data may contain fruitful information about important types of problem – e.g. national security and intelligence, marketing, and medical informatics.

It goes without saying that information is not knowledge. To build (as it were, and as we hope) a cathedral from information-bricks, different types of Deep Learning algorithms are being invented. In a hierarchical learning process, the

algorithms extract and systematise data, and produce data representations. It is argued that this process of analysing unlabelled, uncategorised, unsupervised data will in fact teach us something valuable. It will provide new insights and knowledge.

But Big Data and Deep Learning are not free of values. At some point, someone decided to collect the information – and to collect it the way it has been collected, and to collect it rather than some other information. We can't collect everything. Also the learning mechanisms are value-instilled, at each and every level of the hierarchical process. Someone preferred *this* disambiguated specification of how to solve the problem over *that* one. For a long time now, philosophers of science, Johannes Persson being one of them, have discussed these issues in depth.

In other words, then, Big Data and Deep Learning are not free of values – and so the question "Whose preferences and values?" is inescapable.

Missing the truth

An autonomous system needs a way to update the information it has. Over time, it needs to accumulate proven experience.

In a recent paper, Zalán Gyenis and Miklós Rédei have presented a most interesting result. They study the properties of one of the best-known ways of learning by experience, Bayesian learning. Bayesian learning is a way to

update the probability for a hypothesis via conditionalisation when new evidence becomes available. What Gyenis and Rédei show is that, given a probability measure representing the agent's information and knowledge, there exist probability measures that cannot be learned by the agent from any evidence, or any proven experience (within the given "space" of probabilities). This means that the agent's knowledge and information, represented as a probability measure, might actually prevent the agent from learning the "true" probability, i.e. the probability measure that the agent really wants to learn because it reflects reality.

I don't know if this nice result has any practical consequences, but one thing it definitely tells us is that design matters. The learning of algorithms is a matter of design. A Bayesian learning algorithm comes at a cost. But how big that cost is will depend on the situation, on what is at stake when the updated probability measure is used qua basis for action.

Combining beliefs and values

In "le meilleur des mondes possibles" we are equipped with knowledge and information capable of being represented by a unique probability measure and preferences and values that can be represented by a utility function defined up to an affine transformation. Although probabilities and utilities make no decisions, it can be argued that in the best of all possible worlds a perfectly rational agent maximises

expected utility: she ought to do it and she does it. But the stability of our knowledge influences our decision-making. There are situations in which there are important differences of degree in our knowledge, information or understanding of the various factors underlying our decisions – a difference in ignorance that cannot be mirrored by a unique probability measure. In some situations both information and preferences are unfixed or unreliable. Here we need theories that help us to represent unreliable or indeterminate beliefs and imprecise values. We are obliged to introduce far more complex decision procedures. For simple mathematical reasons "maximise expected utility" is no longer an available option. We must invent generalised theories and decision rules as a basis for action. Ideally these will deliver the same recommendation. They do not. And this is a big problem.

Years ago I was asked if generalised Bayesian decision models could help fighter pilots take decisions in critical situations. Here, I thought, we have theories that are very good at representing the information we find in complex situations. But since these theories use different decision rules, we have a problem. In theory it is not difficult to construct a situation in which the one theory says "fire", another "don't fire", and a third "no idea what you should do". This sounds like a joke, but it is a simple mathematical fact.

This is, of course, as big a problem in a healthcare situation, regardless of whether we are thinking of robotic

surgeons or automated nurses. One way out of the dilemma is to pretend that we have more precise knowledge and information, and more robust preferences and values, than we really have. But idealising in this kind of context comes close to lying. And it doesn't help much.

My point is that the design of autonomous agents, or robots, is very much a question of values at each and every level of the design process. Lurking behind the technical issues at all levels of the process, the stubborn question Whose preferences? Whose values? remains.

Mellor's argument

"How much of the mind is a computer?", D. H. Mellor asks. "Not very much" is his answer. Mellor argues against computational psychologists who seem to believe most, if not all, mental processes are computational. The core of his argument is that while information is true or false, attitudes are not. We compute when we make inferences. Inferences deliver belief; they aim for truth; attitudes most often do not. "Desires, hopes and fears do not embody propositions as true" (p. 79).

Mellor warns us that his question is not "How much of a mind do computers have?" That is another story. Both questions, however, are interesting when we are discussing AI and machine learning, self-driving cars and robotics. One thing that makes us what we are is our preferences and values, our desires, hopes and fears. Do we want the robots,

the autonomous systems, to have a mind in exactly the same sense that we – you and I – have minds? Can we design them that way? If so, they will still not take decisions in the way a rational agent does. There is a question, of course, about how the non-truth tracking bits of the mind, not being beliefs, are to be programmed. Robots need values and preferences in order to be autonomous. But again, whose preferences and whose values? And could they ever, in a meaningful way, be put into the system in such a way that they become the autonomous system's own preferences and values? As elements added non-consensually to the bot, in what sense could they be the bot's own values and preferences?

The Swedish Patient Safety Act requires healthcare professionals to perform their work in accordance with science and proven experience. If they do not, they can be held accountable under penal law. Bad decisions are blameworthy where they are not grounded in science and proven experience. But who is to be held responsible for the actions of robotic doctors and nurses, truly autonomous non-human agents? Who is responsible for their values?

References

Gyenis, Z. Rédei, M. (2017). General properties of Bayesian learning as statistical inference determined by conditional expectations. The Review of Symbolic Logic, 1–37.

Mellor, D. H. (1991). How Much of the mind is a computer? In

Matters of Metaphysics, Cambridge University Press, 61–81.

Santoni de Sio, F., van den Hoven, J. (forthcoming). Meaningful human control over autonomous systems: A philosophical analysis.

van den Hoven, J., Miller, S., Pogge, T. (eds). (2017). Designing in Ethics. Cambridge University Press.

Vanderbilt, T. (2017). Autonomous Cars: How Safe Is Safe Enough? Car and Driver.

Some rather rational reflections on the irrationality of reflection

ROBIN STENWALL

Where possible, we want our decisions to be based on credences founded on scientific evidence and practical experience. But how should we treat beliefs that our degree of belief in a hypothesis will have a certain value in the future? Should your current expectations of your epistemic future self always constrain your belief in the hypothesis, or should there be a restriction on the set of beliefs to which such a principle applies? Suppose, for example, that your degree of belief in a hypothesis (e.g. that surgery X is the best treatment for patient S) is currently 0.9, based on the evidence at hand. Suppose furthermore that you have reason to think that this value will fall to 0.5 next week. The question then is whether it follows from this that your credence in the hypothesis should be 0.5 today, or whether the value should depend on the case at hand. According to Bas van Fraassen (1984), rational agents must always treat their

future selves as experts. Let p be your current credence function and p_t your credence function at some future time t. The Reflection Principle says that if you are rational, then:

(RP) for any H and t, $p(H \mid p_t(H) = r) = r$.[1]

In other words, if you are rational and think that you will assign r to H at t, you should already assign r to H. This looks weird, for we are certainly not rationally required to have this amount of confidence in the credence functions of others. Van Fraassen agrees, stating that (RP) looks "prima facie quite unacceptable" (van Fraassen 1984: 236). However, he insists that reflections on the role of first-person reports of subjective probability should remove any hesitation we might have in accepting the principle. Performative locutions like 'I believe that H to degree r', according to van Fraassen, should not be thought of as descriptive reports of one's psychological state, but rather as undertakings of commitments similar to that made in 'I promise to ϕ'. It is our bias towards descriptivism that makes us think that (RP) is counterintuitive. For if subjective probability judg-

1. van Fraassen thinks that a type of Dutch Book called a Dutch Strategy can be made against anyone who violates this principle. The details of the argument do not matter for present purposes. For a presentation of the diachronic Dutch Book cases that van Fraassen considers, see his (1984). For an argument designed to show why such cases (*pace* van Fraassen) do not support (RP), see Christensen (1991).

ments were nothing but autobiographical reports, there would be no reason for me to think that my subjective belief-states are more reliable than those of others. But such judgments are not descriptions: they're epistemic commitments. And my status as a person of integrity who makes judgments about my own degrees of belief requires that I stand behind my epistemic avowals (van Fraassen 1984: 254). Thus if the rationality involved in belief-formation is analogous to that in sincerely made promises, we should take (RP) to be an essential component of rationality.

I think the principle fails regardless of whether we accept van Fraassen's pragmatic line of reasoning. In fact, I think that there are cases when a violation of (RP) is not only rationally permissible, but mandatory. To borrow an example from David Christensen (1991: 234–35), suppose there is a drug that causes its users to have a credence of 0.99 in the proposition that they can fly. Now, suppose an agent is currently convinced that tomorrow when they take the drug it will make them have a credence of 0.99 that they can fly. According to (RP), this requires them to adopt that unreasonable value already today. Surely this is ridiculous if anything is. The only rational thing for them to do would be to violate (RP) and significantly reduce the risk of jumping to a certain death.

Of course, the above example is quite extreme. One could plausibly argue that any agent who takes such a drug should be deemed irrational (Christensen denies this) or that the

agent's future drugged state is too alien to be taken into account (cf. Jeffrey 1988: 233). However, I think that much more mundane circumstances warrant a violation of (RP). In fact, I think that if we expect our credences to be based on scientific evidence and practical experience, it would sometimes be a mistake to obey (RP). To see this, suppose I'm currently fairly confident, given the evidence, that surgery X is the best type of procedure for S. Say, I believe that hypothesis to a degree of 0.9. But assume I'm also a somewhat squeamish surgeon whose self-confidence plummets at the sight of blood. As an agent with a strong sense of self-knowledge I come to think that on the day of the surgery my credence in the hypothesis will drop to 0.1. According to (RP) it would be rational to assign that value to the hypothesis already today. But again, the only rational thing here would be to look past one's disposition to assign a low probability function to a hypothesis with an abundance of evidence to support it. In fact, one would expect a rational agent to take precautionary measures to avoid acting on one's degree of belief.

To this one might object that the principle is meant to apply only under conditions where the agent has rational reasons for their credence assignments, and that an assignment based on a phobia is hardly rational. Of course, I could make up a story here about our squeamish surgeon having a heart-condition that kicks in when he or she is confident under stress, and thus, that it would in fact be

rational for me to be less confident in the hypothesis on the day of the surgery. But I will not. There is a much easier way to refute (RP).

Bayesian probability theory does not require a uniquely rational subjective probability for every set of evidence. So, suppose I'm epistemically rational (given the set of evidence I have) in assigning different probabilities r_1 and r_2 to the hypothesis *H* that X is the best surgery for S.[2] Suppose furthermore that I have reasons for having a credence of r_1 in *H* today and a credence of r_2 in *H* tomorrow. Such reasons are often non-epistemic. It might be that I'm suffering from hubris and that my CBT therapist asked me to curb my confidence in the surgery, or perhaps I'm planning to charm a colleague who is fond of low-self-esteemed men. Rational reasons like these are far from rare and pose a serious enough threat to (RP). But I think we can do better. For suppose that my plan for tomorrow is to *test* the hypothesis by changing my credence. Perhaps the test allows me to see things from a different point of view, or makes me aware of some of the consequences of *H* that I wouldn't otherwise notice. Either way, my reason for changing between the assignments would be *epistemic*. But if this is correct, then I'm epistemically rational in having a credence of r_1 today

2. In good Bayesian fashion I'm prepared to bet on my credences so that if I believe that *H* to degree *r*, then I'm willing to accept a bet that pays 1 if *H*, nothing otherwise, and which costs *r* units of money. I'm thus assuming that we are dealing with genuine degrees of belief here, not just make believe.

and planning to have a credence of r_2 tomorrow (let us assume that there's no new relevant evidence coming my way). Yet, of course, my current degree of belief in *H* on the supposition that I plan to have a credence of r_2 tomorrow is r_1 and not r_2—thereby refuting (RP).

References

Christensen, D. (1991). Clever Bookies and Coherent Beliefs, The Philosophical Review 100, 229–47.

Jeffrey, R. (1988). Conditioning, Kinematics and Exchangeability in Skyrms, B. & Harper, W. (eds.), Causation, Chance and Credence, Dordrecht: Kluwer, 221–55.

Van Fraassen, B. C. (1984). Belief and the Will, The Journal of Philosophy 81, 235–56.

Rules, norms, evidence, and proven experience

NIKLAS VAREMAN

Introduction

What place does proven experience have in evidence-based medicine (EBM), and is there a relevant difference between EBM and Science and Proven Experience (VBE)? These are questions that Johannes and other participants in the VBE-programme have recently delved into (Persson et al., 2017). The suggestion in this paper, which touches on the interesting historical development of the VBE concept, is that in a rather natural conception of EBM, proven experience is disqualified as a source of evidence. Of course, what proven experience actually consists in is debatable. I will make it easy for myself here and construe it as a firmly held belief which, although it has undergone some kind of testing, is encircled by epistemic support we describe using terms such as "grounded in practice" and "not subjected to scientific testing". It is not research evidence, and as such it is not evidence of the kind prioritized in some formula-

tions of EBM. Proven experience can be information or expertise, but evidence is what comes from proper research and nothing else. Emphasis on the idea that an activity should be based on science and proven experience can be interpreted as showing that both science and proven experience are important sources of evidence, and in this way VBE is different from EBM, it is argued in the paper by Persson et al. (2017). However, it may perhaps seem somewhat pernickety to say that proven experience is excluded from the evidence focused on in EBM when in fact the practitioner can use his or her proven experience in the actual implementation phase. Is this not using evidence in some sense too?

Maybe we can make it a little clearer what this difference amounts to with the help of Carl Hempel's conception of two set of rules governing scientific reasoning and Ilkka Niiniluoto's account of the notion of a technical norm.

Rules of confirmation and acceptance

In "Science and Human Values" (Hempel, 1965) Hempel discusses two sets of rules: *rules of confirmation* and *rules of acceptance*. Rules of confirmation govern what is to be counted as confirmatory and disconfirmatory evidence of a certain hypothesis under investigation. Rules of acceptance state what has to be in place in order to accept or reject a hypothesis, i.e. how much evidence, and of what quality, is needed if we are to accept or reject a hypothesis. The

question when to accept or reject a hypothesis rests on the risk that a false hypothesis will be accepted, or the risk that a true hypothesis will be rejected. Hempel calls these risks "inductive". The rules of acceptance then decide what level of inductive risk we can accept. So, then, what is meant by "acceptance" ? Here we will settle for the idea that acceptance is the decision to use a hypothesis as a basis for a decision. So, for instance, if as a doctor you decide to treat your patient, suffering from a headache, with Aspirin, you have accepted the hypothesis that Aspirin is effective in treating headache.

The rules of confirmation in EBM are quite clearly formulated both by way of evidence hierarchies that have for some time been the basis of systematic reviews of medical literature (CEBM, 2009) and by the Grading of Recommendations, Assessment, Development and Evaluations (GRADE) framework for evidence assessment. The latter describes how we are to rate evidence, up or down, by assessing factors including: study design (inherently, randomized controlled studies are of higher quality than observational studies and case studies, etc.), study limitations, inconsistency, indirectness, imprecision, publication bias, magnitude of effect, dose-response gradient and whether confounding factors would lessen an effect, or suggest a spurious effect if no effect was found. These are rules of confirmation; they decide what is to be regarded as confirmatory or disconfirmatory evidence, and to what degree. The rules exclude

proven experience. Expert opinion, a species of proven experience one might argue, is at the absolute bottom of the evidence hierarchy set out by the Centre for Evidence-Based Medicine (CEBM, 2009) and is not even mentioned in GRADE.

On the other hand, clinical expertise is mentioned in EBM as an important part of implementing evidence-based methods in practice. It plays a significant role in the acceptance of a method. In EBM, then, proven experience, or clinical expertise, is allowed in by the rules of acceptance, it seems. To what degree this is actually the case can perhaps be debated. In GRADE there are rules of acceptance as well as rules of confirmation, and the former are based on risk-benefit analyses where one decides – on the basis of thresholds describing how many patients need to be treated in order to achieve one successful effect – whether the benefits outweigh the risks. A strength of recommendation is set using this measure. If this recommendation becomes, at the policy level, a guideline for use, then proven experience of practitioners is not part of the rules of acceptance. But the values of the people making GRADE recommendations do inform those rules. So a rigid use of GRADE could exclude proven experience from the medical decision making altogether. We will, however, follow the intent of EBM to let the rules of acceptance allow clinical expertise in the decision making.

The rules of confirmation relating to VBE do, one may

argue, allow proven experience to provide evidence confirming or disconfirming hypotheses – together, of course, with evidence from EBM.

Evidence and technical norms

Hempel notes that the question whether a hypothesis should be accepted or rejected depends on the degree to which it reaches a goal of some sort. This goal can, in practical circumstances, be economic, or technological, or – as in our case here – related to health. In pure science things are, perhaps, less clear, but Hempel argues "that the standards governing the inductive procedures of pure science reflect the objective of obtaining a certain goal, which might be described somewhat vaguely as the attainment of an increasingly reliable, extensive, and theoretically systematized body of information about the world" (Hempel, 1965). These standards could be different. Presumably, the goal could be aesthetic in character, and then the rules of confirmation and acceptance would be different too, certainly. Hempel continues: "the standards of procedure must in each case be formed in consideration of the goals to be attained; their justification must be relative to those goals and must, in this sense, presuppose them" (ibid).

So, one may ask, what are the goals, or values, that are presupposed by the exclusion of proven experience from the realm of evidence?

The rules of acceptance in GRADE, as described above, suggest that the goal of EBM is to provide *technical norms* (see Niiniluoto, 1993, for description of G. H. von Wright's notion). These are statements of the form "If you want A, and you believe that you are in situation B, then you ought to do X". In the case of medicine they take a form such as "If you want to make a certain proportion of patients, P, suffering from disease D, healthy, you ought to treat these using method M".

The technical norm might be true if method M actually is effective in treating disease D and consequently restores health in P – that is, if M causes P to be healthy if P has disease D. This can be established either "from above" (Niiniluoto, 1993) by derivation from general causal statements, laws, established from (pure) science. Or it can be done "'from below' by building up a simplified model of the situation, using trial-and-error procedures and experimental tests to investigate the dependences between the most important variables, and trying to find the optimal methods of producing the desired effects. When the result is expressed as a general rule, a technical norm with some empirical support is obtained" (Niiniluoto, 1993). Simplifying, and adapting the suggestion to the present context, we can say, then, that research evidence supports from above and proven experience supports from below. EBM is only satisfied with evidence from above, while VBE accepts support from both above and below.

All this may seem quite as it should be, since EBM is about deciding how to treat patients on the basis of what the science says is effective treatment. But a technical norm can get adequate support also from below, so why not accept this way of confirming the norm if it is the validity of the norm we are after rather than the truth of a causal claim? The rules of acceptability suggest that the goal of EBM is promoting health; it is not about truth or a "theoretically systematized body of information". With a large enough effect, highly uncertain methods can be accepted if the effect, expressed with its uncertainty, is above a set clinical threshold. So, why should the rules of confirmation aim at the (possibly) higher goal of truth when the rules of acceptance settle for the goal of promoting health?

References

CEBM. (2009). Oxford Centre for Evidence-based Medicine – Levels of Evidence (March 2009). Accessed 2017 November 17. http://www.cebm.net/oxford-centre-evidence-based-medicine-levels-evidence-march-2009/.

Hempel, C. (1965). Science and human values, in Aspects of Scientific Explanation, And Other Essays in the Philosophy of Science, The free press, New York.

Niiniluoto, I. (1993). The aim and structure of applied research, Erkenntnis, 38:1–21.

Persson, J., Vareman, N., Wallin, A., Wahlberg, L., Sahlin, N-E. (2017, forthcoming). Science and proven experience: a Swedish variety of evidence based medicine? Journal of Risk Research.

Helheten och delarna

Kan ”vetenskap och beprövad erfarenhet” alltid reduceras till ”vetenskap”, ”och” och ”beprövad erfarenhet”?

LENA WAHLBERG

Inledning

För att förstå vad ”vetenskap och beprövad erfarenhet” betyder räcker det inte att veta vad vetenskap är och vad beprövad erfarenhet är. Vi behöver också veta hur de två komponenterna förhåller sig till varandra: vad konjunktionen – ”och”:et – i ”vetenskap *och* beprövad erfarenhet” egentligen står för. Johannes Persson m.fl. har föreslagit att konjunktionen mellan ”vetenskap” och ”beprövad erfarenhet” kan läsas ”och/eller-men-inte-i-strid-med-någon”, men också visat att förståelsen av ”beprövad erfarenhet” inte är entydig utan varierar mellan olika användare. En naturlig följdfråga är om konjunktionens innebörd i själva verket varierar med förståelsen av ”beprövad erfarenhet”. Inte minst när förståelsen av ”beprövad erfarenhet” ligger långt från förståelsen av ”vetenskap” kan förmodas att den

ena inte utan vidare kan ersätta den andra. I denna text kommer jag att presentera några tankar om detta. Min tentativa slutsats är att innebörden av konjunktionen varierar med förståelsen av ”beprövad erfarenhet”, och att variationen beror på hur sammanhanget påverkar innebörden av *hela* begreppet ”vetenskap och beprövad erfarenhet”. En analys av delarna är viktig för att förstå helhetens spännvidd, men inte tillräcklig för att bestämma helhetens och delarnas innebörd i ett konkret fall.

”Och”

I artikeln ”Hur förstå ’och’ i ’vetenskap och beprövad erfarenhet’” frågar sig Johannes Persson, Sten Anttila och Nils-Eric Sahlin om ett krav på överenstämmelse med vetenskap och beprövad erfarenhet ska tolkas som ett krav på överensstämmelse med *både* vetenskap *och* beprövad erfarenhet, eller om det kanske räcker med överensstämmelse med vetenskap *eller* beprövad erfarenhet. Författarna konstaterar att tolkningen *både och* blir svår att upprätthålla och kan vara onödigt sträng i de fall då vetenskap och beprövad erfarenhet är i otakt och vi bara har det ena, men att tolkningen *och/eller* å andra sidan blir alltför svag i de situationer där vetenskap och beprövad erfarenhet är motstridiga. Mot denna bakgrund presenterar författarna i stället den alternativa tolkningen ”och/eller-men-inte-i-strid-med-någon”. Att X är i överensstämmelse med vetenskap och beprövad erfarenhet betyder med denna tolkning att X

är i överensstämmelse med vetenskap och/eller med beprövad erfarenhet och att X inte strider mot någon av dessa. Denna tolkning säger inte hur mycket vetenskap eller hur mycket beprövad erfarenhet som krävs, och inte heller vad vetenskap eller beprövad erfarenhet är, men den föreslår ett kvalitativt minimikrav för att kravet ska vara uppfyllt.

”Beprövad erfarenhet”

I artikeln ”Vår erfarenhet av beprövad erfarenhet” undersöker Johannes Persson och jag användningen av ”beprövad erfarenhet” i texter publicerade i Läkartidningen. I artikeln identifieras följande sex dimensioner hos begreppet: (1) prövningens allvar; (2) praktiken som ursprung; (3) praktiken som prövningsmekanism; (4) praktiken som evidens; (5) erfarenhetens utbredning: personen; och (6) erfarenhetens utbredning: kollektivet. Beroende på var tonvikten placeras framträder olika begreppsprofiler och därmed olika betydelser av ”beprövad erfarenhet”. Som Johannes Persson visat återkommer flera av dessa dimensioner – med delvis annan tonvikt – i andra sammanhang, inte minst i skolans värld. I den kommande diskussionen kommer jag för enkelhetens skull att utgå från tre olika förståelser av beprövad erfarenhet, som alla ger uttryck för en eller flera av dessa dimensioner. Jag kommer att kalla de tre förståelserna *be-prövning, be-praxis* och *be-träning*.

Be-prövning, be-praxis och be-träning

Enligt en vanlig förståelse av beprövad erfarenhet handlar beprövad erfarenhet om *prövning*. Att det finns beprövad erfarenhet av X innebär med denna förståelse (nedan betecknad *be-prövning*) att X har prövats i praktiken, och att detta gett information om vad som händer när X används. Be-prövning betonar dimensionerna 1, 3 och 4 ovan. Den beprövade erfarenheten kan visa om X fungerar.

> Mer än 10 års beprövad erfarenhet visar att de flesta vårtor går bort om man behandlar med VårtFri regelbundet en gång i veckan.

Beprövad erfarenhet i betydelsen be-prövning har delvis samma funktion som vetenskapen (forskning): att ge evidens för en behandlings eller åtgärds effektivitet. Likheten blir kanske särskilt tydlig när begreppen används parallellt, vilket är vanligt också i utom-medicinska sammanhang. Här ett exempel från Brottsförebyggande rådet:

> Svensk polis bör givetvis sträva efter att i så hög utsträckning som möjligt i första hand genomföra sådana åtgärder som genom forskning eller beprövad erfarenhet har visat sig fungera.

Enligt en tillsynes väsensskild förståelse handlar beprövad erfarenhet i stället om vad som är *praxis*. Att det finns

beprövad erfarenhet (nedan betecknad *be-praxis*) av X innebär att X är en vedertagen metod (eller uppfattning) – X är vad man i praktiken *gör* (eller anser). Be-praxis, som betonar den sjätte dimensionen, är enligt min erfarenhet en ganska vanlig förståelse av beprövad erfarenhet i juridiska sammanhang. Det uttalande Johannes Persson och jag använde för att illustrera den sjätte dimensionen i Läkartidningen kom också mycket riktigt från en jurist på Socialstyrelsen:

> Beprövad erfarenhet innebär sådana metoder som används inom vården och anses vara verksamma. Det som läkarkollektivet anser vara en inarbetad praxis kan innefattas här.

Frågan om utbredningen i ett kollektiv är inte nödvändigtvis rent kvantitativ i förhållande till läkarkollektivet i dess helhet, utan kan avse en på ett relevant sätt avgränsad grupp. Så här står det till exempel i propositionen till patientskadelagen:

> Bedömningen [av om behandlingsmetoden varit i överensstämmelse med vetenskap och beprövad erfarenhet] sker med utgångspunkt i *den erfarne specialistens* kunskap vid tidpunkten för behandlingen. Om man då finner att den valda behandlingsmetoden *inte är en i praxis vedertagen metod*, utges ersättning för skadan. (min kursivering)

Be-praxis handlar inte direkt om prövningen. Förvisso kan det som är prövat också vara praxis (och vice versa), men det behöver inte vara så: praxis kan ha sin grund i tradition, och kanske ibland förklaras av bekvämlighet och en tröghet att ta till sig nya rön.

Den tredje förståelsen jag vill ta upp betonar den femte dimensionen. Här handlar beprövad erfarenhet om *träning*. Enligt denna förståelse är den beprövade erfarenheten (nedan betecknad *be-träning*) inte bara kopplad till en viss behandlingsmetod, X, utan också till någon, en ”person”, Y, till exempel en fysisk individ eller en forskargrupp.

Be-träning är en vanlig förståelse av beprövad erfarenhet när intresset i första hand avser *personens* – inte *behandlingsmetodens* – kvaliteter: ”Y är en person som har beprövad erfarenhet av X”. Inte minst i rättsliga sammanhang används be-träning emellertid också som ett mått på behandlingsmetodens kvaliteter: ”X är en behandlingsmetod som ett Y har beprövad erfarenhet av”. I rättsfallet RÅ 2004 ref. 41, som rörde ersättning för gränsöverskridande vård, konstaterade till exempel Högsta förvaltningsdomstolen att den behandling som patienten fått i Kiel:

> under flera års tid använts vid universitetskliniken i Kiel för behandling av ett antal patienter, av vilka flera lidit av [den aktuella sjukdomen]

Trots att det vetenskapliga stödet föreföll närmast obefintligt beviljade Högsta förvaltningsdomstolen ersättning. Med stöd av detta avgörande fann Förvaltningsrätten i Stockholm i en serie senare fall att be-träning i princip var en tillräcklig förutsättning för att en behandling skulle anses överensstämma med vetenskap och beprövad erfarenhet (bedömningen i dessa fall avsåg hypertermibehandling av cancer och stod sig inte i kammarrätten).

Medan be-praxis (dimension 6) betonar utbredningen av erfarenheten i ett kollektiv betonar be-träning (dimension 5) utbredningen av erfarenheten hos en person. Även om också be-prövning kan användas som ett mått på behandlingsmetodens kvaliteter tycks den, till skillnad från be-praxis, vara bunden till personen. Kopplingen mellan behandlingsmetoden, erfarenheten och personen blir kanske ännu tydligare i följande uttalande av en klinikchef, hämtat från ett färskt rättsfall om behandling av generell hyperhidros:

> Denna behandling [av generell hyperhidros] hade ett begränsat vetenskapligt underlag och utfördes av en läkare med mångårig erfarenhet av svettbehandling, och därmed i överensstämmelse med väl beprövad erfarenhet. Denna specifika läkarkompetens kommer inte finnas tillgänglig som tidigare och sjukhuset kommer därför inte längre att utföra behandling av generell hyperhidros.

Helheten och delarna

I sin artikel om konjunktionen tycks Johannes Persson, Sten Anttila och Nils-Eric Sahlin i första hand förstå beprövad erfarenhet som *be-prövning*. Liksom tolkningarna ”både och” och ”och/eller” behandlar tolkningen ”och/eller-men-inte-i-strid-med-någon” den beprövade erfarenheten och vetenskapen *symmetriskt*. En symmetrisk behandling förefaller rimlig när den beprövade erfarenheten förstås som be-prövning, och liksom vetenskapen kan ge evidens för en behandlingsmetods effektivitet. Som argument mot ”både och”-tolkningen anför författarna att vetenskapen och den beprövade erfarenheten ibland är i otakt, vilket passar väl med just be-prövning-förståelsen. Artikelns argumentation mot en ren ”och/eller”-tolkning stödjer sig främst på den vardagsspråkliga innebörden av ”överensstämmelse”: det som strider mot vetenskap eller beprövad erfarenhet kan inte överensstämma med vetenskap och beprövad erfarenhet. Även i denna del tycks det dock gå att hitta materiella skäl för ”och/eller-men-inte-i-strid-med-någon”-tolkningen. Vi kan till exempel knappast ignorera beprövad erfarenhet som visar att en behandlingsmetod leder till oacceptabla skador, även om vetenskapen skulle visa att behandlingen också har positiva effekter. Om vi å andra sidan antar att den beprövade erfarenheten och vetenskapen kan ge kunskap om olika effekter, verkar det problematiskt att underkänna en behandlingsmetod som enligt den beprövade erfarenheten har goda effekter, med hänvisning till att dessa

effekter inte kunnat identifieras i några vetenskapliga studier (och vice versa) – i vart fall om effekterna är av ett sådant slag att de av moraliska, kunskapsteoretiska eller andra skäl inte kan studeras vetenskapligt. Om bedömningen tar hänsyn till de olika kunskapstypernas räckvidd – vilket jag uppfattar att författarna är öppna för – kan dock ”och/eller-men-inte-i-strid-med-någon”-tolkningen rättfärdigas även här.

Fungerar ”och/eller-men-inte-i-strid-med-någon”- tolkningen lika väl när beprövad erfarenhet förstås som be-praxis eller be-träning? Utan att fördjupa oss i den knepiga frågan vad *i strid* med be-praxis och be-träning betyder, kan vi konstatera att det finns fall där också be-praxis och be-träning tillerkänts stor betydelse. I de uttalanden om be-träning som citerades i föregående avsnitt ansågs i princip det faktum att en aktör hade omfattande erfarenhet av en viss behandlingsmetod (Kiel-protokollet, hypertermi-behandling respektive behandling av hyperhidros) vara tillräckligt för att metoden skulle anses acceptabel och i överensstämmelse med vetenskap och beprövad erfarenhet. I andra fall (till exempel Socialstyrelsens yttrande i kammar-rätten i fallen om hypertermibehandling) har den omstän-digheten att det saknas be-praxis av behandlingsmetoden (att ”hypertermi inte är en rutinmässigt och kliniskt invänd-ningsfri rutinbehandling”) anförts som tungt vägande skäl för att metoden inte överensstämmer med vetenskap och beprövad erfarenhet. Det *förekommer* alltså att också

be-praxis och be-träning tillerkänns samma vikt som vetenskaplig evidens. Men är *skälen* för en symmetrisk behandling i dessa fall lika goda som när det gäller be-prövning?

För att besvara denna fråga behöver vi ta ställning till om det är delarna (”vetenskap”, ”och” och ”beprövad erfarenhet”) som bestämmer helhetens innebörd eller tvärtom. Själv misstänker jag att vi behöver ta hänsyn till helheten, och inte minst till hur sammanhanget påverkar helhetens innebörd, för att fullt ut kunna bedöma delarnas betydelse i ett konkret fall. Vissa användningar av begreppet ”vetenskap och beprövad erfarenhet” tyder rentav på att helhetens innebörd inte alltid kan analyseras i termer av delarna. I propositionen om förbud mot omskärelse av kvinnor, står till exempel att läsa att ”Socialstyrelsen har uttalat att kvinnlig omskärelse i alla former enligt styrelsens uppfattning står i strid mot vetenskap och beprövad erfarenhet”. Av Socialstyrelsens föreskrifter om livsuppehållande behandling framgår vidare att livsuppehållande behandling inte alltid är förenlig med vetenskap och beprövad erfarenhet. Vilken roll spelar vetenskapen och den beprövade erfarenheten här? Är det rentav så att användningen av begreppet ”vetenskap och beprövad erfarenhet” i dessa fall ger uttryck för vad som är etiskt acceptabelt, utan att detta nödvändigtvis låter sig förklaras med hänvisning till vad vetenskapen och den beprövade erfarenheten säger? Kanske betyder ”vetenskap och beprövad erfarenhet” ibland något mer, eller i vart fall något annat, än vetenskap och beprövad erfarenhet?

I andra användningar tycks den ena komponenten helt dominerande. I vissa fall förefaller ”vetenskap och beprövad erfarenhet” helt enkelt betyda vedertagen medicinsk praxis. (Citatet från propositionen till patientskadelagen ovan är möjligen ett exempel på detta.) Med den innebörden av helheten blir det förstås naturligt att tillerkänna be-praxis avgörande betydelse medan den vetenskapliga evidensen får stå tillbaka: en behandlingsmetod överensstämmer med vetenskap och beprövad erfarenhet om och endast om den är vedertagen medicinsk praxis. Om ”vetenskap och beprövad erfarenhet” i stället betyder ”tillräcklig evidens för en behandlingsmetods säkerhet och effektivitet” blir läget ett annat. Be-praxisens värde torde med denna betydelse bero på den förmodade korrelationen mellan be-praxis och be-prövning, och därtill utan vidare trumfas av konstaterad be-prövning och vetenskaplig evidens.

De nu diskuterade exemplen tyder på att innebörden av konjunktionen, ”och”:et, i begreppet ”vetenskap och beprövad erfarenhet” varierar med såväl innebörden av ”beprövad erfarenhet” som med innebörden av hela begreppet ”vetenskap och beprövad erfarenhet”. Mot detta skulle kunna invändas att det finns ”sanna” förståelser av komponenterna ”vetenskap”, ”och”, och ”beprövad erfarenhet”, och att endast en helhet som härbärgerar dessa sanna förståelser av delarna utgör en rimlig förståelse av helheten. Det kan förvisso diskuteras om alla förekommande förståelser av ”vetenskap och beprövad erfarenhet” är rimliga.

Jag förmodar att så inte är fallet men tror samtidigt att bedömningen av vad som är en rimlig förståelse måste ske i ljuset av vad som bäst tjänar *hela* begreppets funktion i ett visst sammanhang. I juridiken, till exempel, kommer funktionen hos begreppet ”vetenskap och beprövad erfarenhet” att variera med de rättsregler i vilka det förekommer: I vissa regler är begreppet en allmän ledstjärna för vårdarbetet, i andra regler definierar det patienters rätt att välja mellan olika behandlingsmetoder och i åter andra avgränsar det den enskilde personalens ansvar eller patientens rätt till ersättning för vårdskada. Vad som är en rimlig förståelse av begreppet beror på de olika överväganden som ligger till grund för dessa rättsregler, vilka i sin tur får betydelse för vad som i ett visst sammanhang är en rimlig förståelse av begreppets delar, inklusive konjunktionen. En analys av delarna är viktig för att se helhetens spännvidd men räcker sannolikt inte för att bedöma vare sig helhetens eller delarnas innebörd i ett konkret fall. En konsekvens av detta är att relevansen av sakkunnigutlåtande av typen ”behandlingsmetoden är inte en rutinmässigt och kliniskt invändningsfri rutinbehandling” inte utan vidare kan tas för given, utan måste bedömas i ljuset av vad som i sammanhanget är en rimlig innebörd av hela begreppet ”vetenskap och beprövad erfarenhet”.

Referenser

Lindahl, E., Lindström, P. och Svensson, D. (2004). Effektiva polisiära åtgärder mot brott: En sammanställning av forskning och beprövad erfarenhet, BRÅ 2004.

Persson J., Anttila, S. Sahlin, N-E., Hur förstå 'och' i 'vetenskap och beprövad erfarenhet', utkommande i Filosofisk tidskrift.

Persson J. (2017). Är vetenskaplig grund och beprövad erfarenhet i skolan samma sak som vetenskap och beprövad erfarenhet i hälso- och sjukvård? VBE Skola, Lund 2017.

Persson J., Wahlberg, L. (2015). Vår erfarenhet av beprövad erfarenhet: Några begreppsprofiler och ett verktyg för precisering, Läkartidningen 49/2015.

Kammarrätten i Stockholm, domar meddelade 2014-12-03 i mål 2418-13, 2419-13, 2420-13, 3609-13, 3615-13, 3610-13, 3613-13 och 3642-13.

Produktbeskrivning Vårtfri, 30 behandlingar, 5 ml, https://fiorina.se/hudvard/ovrig-hudvard/vartfri-30-behandlingar-5ml/ (hämtad 2017-10-30).

Proposition 1995/96:187 Patientskadelag m.m., s. 33.

Proposition 1981/82:172 om förbud mot omskärelse av kvinnor, s. 9.

RÅ 2004 ref 41. SOSFS 2011:7, Socialstyrelsens föreskrifter och allmänna råd om livsuppehållande behandling, 3 kap.

Science, proven experience and good sense

ANNIKA WALLIN

Science is a wonderful thing. Done right, it fulfils high standards of reliability and replicability and it will, moreover, give us an idea of how things relate: be it causally or in other ways. Therefore, it is not surprising that we want decisions to be based on science and scientific evidence. This is true for, among other things, forestry management, climate adaptation, education, policy and medical care.

But what does it mean to base decisions on scientific evidence? This seems to vary with the domain. In principle, what we have is often two types of knowledge. The first is evidence that some things do or do not *work* in particular contexts. The second is ideas and evidence about *why* things work or do not do so in this context. Process and outcome, to put it briefly.

One of the best ways of acquiring knowledge about what works or does not work is through randomized controlled

trials. In these, two conditions are kept equal except for one critical factor. Individuals or other types of entity are randomly assigned to each, and the outcomes are compared. If they differ, this must be due to the critical factor – to the intervention. Randomized controlled trials are often described as a gold standard for research. The problem with them is that this way of doing research may lead to a position where we know *that* the intervention (the critical factor) produced a difference, but we do not know *why*. When we do not know why we also do not know how successfully the finding will generalize. Of course, more controlled trials can be performed, systematically testing aspects of the set-up, but we will inevitably end up at a point where we have to make a leap of faith as to whether or not the results are likely to hold in another situation.

In some cases, this leap of faith has to be made almost immediately. When selecting policies, policy makers often have to rely on evidence about what outcomes a particular policy produced in other areas and settings – say, in another country. The original data may, at least in principle, be of very good quality, and even obtained through a randomized controlled trial. Nevertheless, it will be difficult to estimate whether the differences produced by the intervention depend on country-specific details (Persson et al., forthcoming). The same is also true on a smaller scale, education being a case in point (Persson, 2017). The fact that a particular procedure and intervention worked in one school

doesn't necessarily imply that it will work in other schools. Staff and students will differ, as will the schools' organizational set-ups. Two educational systems, two societies or two schools have to be *sufficiently* similar in relevant aspects for a generalization to be sound.

This leads us to the difficult question: What does it mean to be sufficiently similar? The question is almost impossible to answer if we do not at least have some ideas about *why* the intervention might have produced the effect. When we move between domains (countries, schools, hospital systems) we have to learn more about the process producing the outcome, because it is only when we have ideas about why something works that we generate ideas about what will make it *not* work (Persson & Wallin, 2015).

Sometimes those ideas come easily. Some things are known to warrant immediate caution. If we know, for example, that there are relevant differences between the population of a scientific study and a population we are considering applying the results to, we should be careful. A case in point is the fragile elderly, often with multiple diseases, who are treated in primary care. Most scientific studies in the field of medicine are made on individuals with just one predominant disease or affliction. For good reasons, both with respect to ethics and with respect to the soundness of the study, the elderly with some sort of dementia, or with several severe diseases, and people at the very end of their life, are rarely included in scientific studies. Their absence,

however, is also problematic. A large portion of the people that are given medical care are elderly, have several diseases, have dementia or may be at the end of their lives. As well as disqualifying individuals from taking part in medical studies, these attributes also qualify them to receive medical care.

This is, of course, something that professionals in medicine are aware of. When assessing a patient, they have to judge whether the individual is *different* enough from (or similar enough to) the individuals studied in the randomized controlled trials favoured in evidence-based medicine. For instance, anti-coagulants have historically been less frequently administered to the elderly and the elderly with dementia. This is because although the anti-coagulants may reduce the risk of stroke, they also increase the risk of haemorrhage. Since we know that the elderly and the elderly with dementia are more prone to falling, anti-coagulants have been administered to this group with caution. (It appears, however, that this fear may be exaggerated, and that the benefits of anti-coagulants outweigh its risks also for this group, see SBU, 2014.)

The problems presented by the fragile elderly population can be seen relatively easily, even if they are difficult to deal with. If study populations are not representative, we know that caution has to be exercised when extrapolating from them. In other cases, the obstacles to successful generalization may be more hidden. The advantage of pharma-

ceuticals is that although people differ, they all have bodies, and we have relatively well-established ideas about what may affect the human body's uptake or reaction to particular drugs. More procedural interventions – such as toilet training to handle incontinence – depend more heavily on how specific care facilities are organized as well as the bodily specifics of the people being treated. Here, it is more difficult to fathom what it may be in the organizational set-up that allows, or does not allow, a specific method to be successful.

Unfortunately, this does not mean that "why" questions alone will save us. Above I hinted that human bodies are more or less similar, but unfortunately, they are also very complex. One of the reasons why randomized controlled trials became so popular in medicine was evidence that we cannot conclude that things that should work in theory (where we have some idea about why) will also work in practice. An example given by Holly Andersen (Andersen, 2012) is the prophylactic use of paracetamol when infants are vaccinated. Since vaccinations may give rise to fever, and fever can be brought down with paracetamol, this appears on the surface to be a no-brainer. As it turns out, however, the two interventions interact, and paracetamol should only be administered remedially when an infant does develop a fever in reaction to the vaccination. The reasons for this are complicated, and the general practitioner will not necessarily be conversant with them, or so Andersen

claims. In general, the body is filled with interacting and counteracting processes, and they relate in such complex ways that our knowledge why and our knowledge that can appear to be in conflict with each other. We need both, and we need them simultaneously. This is not an easy task, and it requires something special from the person who is to make decisions based on scientific evidence. That individual needs to have "good sense". Where does this good sense come from? No one knows. It is clear, however, that many ways of reasoning and acquiring knowledge will be necessary.

To borrow again, from Pierre Duhem, just as the body is more than its tissues, and a hospital more than a collection of procedures, more than one method of finding the truth is required to use science well in decision-making. The "perfect form of science could not be obtained except by a very precise separation of the various methods concurring in the discovery of truth. Each of the many faculties that human reason puts into play when it wishes to know more and better would have to play its role, without anything being omitted, without any faculty being overlooked. This perfect equilibrium between the many organs of reason does not occur in any one man. In each of us one faculty is stronger and another weaker. In the conquest of truth the weaker will not contribute as much as it should and the stronger will take on more than its share" (Duhem, 1915). This is why we need a little of all. We need both science and something else – be it good sense, or proven experience.

References

Andersen, H. (2012). Mechanisms: what are they evidence for in evidence-based medicine? Journal of Evaluation in Clinical Practice, 18(5).

Duhem, P. (1915). La science allemande. Paris: Hermann. English translation John Lyon, German Science, La Salle, IL: Open Court, 1991.

Statens beredning for medicinsk utvärdering (2014). Nytta och risk med läkemedel för äldre: perorala antikoagulantia och trombocythämmare.

Persson, J. and co-authors (forthcoming). Harnessing local knowledge for scientific knowledge production – challenges and pitfalls.

Persson, J. (2017). Är vetenskaplig grund och beprövad erfarenhet i skolan samma sak som vetenskap och beprövad erfarenhet i hälso- och sjukvård? VBE Skola, Lund.

Persson, J., Wallin, A. (2015). The (misconceived) distinction between internal and external validity. In Persson, J., Hermerén, G., Sjöstrand, E. (Eds.) Against boredom: 17 essays on ignorance, values, creativity, metaphysics, decision-making, truth, preference, art, processes, Ramsey, ethics, rationality, validity, human ills, science, and eternal life to Nils-Eric Sahlin on the occasion of his 60th birthday. Fri tanke förlag.

About the authors

STEN ANTTILA, Ph. D., Project Leader, Swedish Agency for Health Technology Assessment and Assessment of Social Services, Stockholm

WÄNDI BRUINE DE BRUIN, Professor with University Leadership Chair in Behavioural Decision Making, Leeds University Business School

JOHAN BRÄNNMARK, Associate Professor in Ethics and Political Philosophy at the Department of Global Political Studies, Academy Researcher, financed by The Royal Swedish Academy of Letters, History and Antiquities, Malmö University

ALEX DAVIS, Assistant Professor in the Department of Engineering and Public Policy, Carnegie Mellon University

BARRY DEWITT, Postdoctoral Research Scientist, Department of Engineering and Public Policy, Carnegie Mellon University

BARUCH FISCHHOFF, Howard Heinz University Professor, Department of Engineering and Public Policy, Institute for Politics and Strategy, Carnegie Mellon University

CHARLOTTA LEVAY, Associate Professor, Department of Business Administration, Lund University

NILS-ERIC SAHLIN, Professor, Medical Ethics, Faculty of Medicine, Lund University

ROBIN STENWALL, Senior Lecturer, Theoretical Philosophy, Lund University

NIKLAS VAREMAN, Associate Researcher, Medical Ethics, Lund University

LENA WAHLBERG, Senior Lecturer, Associate Professor, Jurisprudence, Department of Law, Lund University

ANNIKA WALLIN, Associate Professor, Cognitive Science, Lund University