ESSAY REVIEW

Does anyone know the road from a randomized trial to personalized medicine? A review of 'Treating Individuals. From Randomized Trials to Personalised Medicine' Peter M. Rothwell

Elsevier, UK. ISBN-13: 978-0-08-044739-1, pp. 340

Eyal Shahar MD MPH

Professor, Division of Epidemiology and Biostatistics, Mel and Enid Zuckerman College of Public Health, University of Arizona, Tucson, AZ, USA.
E-mail: shahar@email.arizona.edu

Introduction

If you ever decide to write and sell a book, you should be aware that the title carries more weight than the content, and sometimes the title carries most of the weight. *Treating Individuals: From Randomised Trials to Personalised Medicine* (edited by Peter M. Rothwell) was a clever choice of a book title [1]. Sounds like the text you have always wanted to read: words of scientific wisdom that will teach you how to weigh the evidence from randomized trials and how to apply the evidence to Mrs Smith in the corner bed in room 376, or to Mr Jones who is about to enter your clinic. The instructions are on the long side – over 300 pages – with about two dozen contributors who 'have not generally seen or commented on other chapters' (pp. x–xi). Consequently, as the editor writes, 'the views expressed are sometimes inconsistent, or even contradictory'. That honest remark is, unfortunately, not the only deficiency of what is supposed to be your friendly teaching tool. I assume, of course, that a book on medical practice is written to teach how to practice medicine and not merely 'to present . . . different perspectives'. (p. xi)

First and foremost, the *Lancet* book looks more like a collection of 20 personal essays of varying quality and relevance than linear progression of chapters from the first to the last. A book should be more than the pieces that make it – more than solicited essays bound in a paper cover. In that respect, the *Lancet* has produced no more than an expanded version of a thematic journal edition, most of which was catered to the rhetoric of the so-called evidence-based medicine movement [2,3]. The difficulty of combining the essays into a coherent book becomes evident as soon as you try to make sense of the Table of Contents. The collection was divided into four sections: Section 1, titled 'Reliable determination of the overall effects of treatments', contains just two essays, one of which – with a 24-word long title – apparently equates reliable determination with systematic reviews. To get the title of section 2, take the title of section 3 ('Is the *overall* trial *result sufficiently* relevant to this patient?') and delete the words I italicized. So, do you think that section 2 is dealing with trials that are *insufficiently* relevant to this patient or with *part* of the results? Not really. Then, you find the fourth and last section titled 'Targeting of treatment in routine practice'. Judging from the title, that section should teach you who to treat in routine practice (in contradistinction to non-routine practice?), and I am still wondering about the difference between that topic and the topics of sections 2 and 3. As you read the book, you can try to decide whether the essays in each section indeed uniquely belong there. In other words, you should decide if you are holding not only a physical object called a book, but also a unified intellectual entity called a book.

This collection of essays claims to teach you a lot in the domain of cause-and-effect, one of the most challenging topics in epidemiology, statistics and philosophy of science. To my mind, a volume on cause-and-effect in medicine should include a thorough discussion of epistemology (theories of knowledge), the various meanings of probability, models of causation, causal parameters, estimators and their desired properties (unbiasedness, efficiency, consistency), *fallible estimates* (versus expected values), effect-measure modification, sources of random variation, clashing schools of statistical thought and more. If you search for these terms in the 'Index', you will find little to nothing. To me, that's a bad sign. That's an indication of text that avoids the real challenges of causal inquiry and causal inference.

Two unanswered questions

So, what is this book all about, besides carrying a good title? The author of the Foreword explains it succinctly:

> Two concepts dominate. First, reliability: how consistently good are the data . . . in terms of inherent quality (otherwise known as internal validity)? Second, relevance: how appropriate are those data to the problem the patient presents (sometimes called external validity)? Reliability and relevance are the only two questions that should matter to the physician . . . (p. xiii)

Reliability and internal validity are distinctly different concepts, at least in the jargon of research methods; whereas, external validity (and its various synonyms [4] might win the first prize for the most commonly abused foggy idea. Terminology aside, I agree that a physician is interested in the answers to two questions: (1) What is the quality of the research on some treatment? and (2) Will that treatment meaningfully help Mrs Smith? Let me offer my answers.

To be able to scrutinize a study thoughtfully, you have to study research design and data analysis in some formal or informal setting. It takes at least 2 years of coursework to establish the foundation of knowledge, and many more years to explore the material in-depth. Although good (and bad) ideas are scattered throughout the essays, the *Lancet* collection is not a serious substitute for textbooks on research methods and biostatistics. It does not systematically teach you the methods of biomedical research, which made me wonder why those, who praise the idol of *systematic reviews*, have failed to produce a *systematic* teaching tool. If you don't trust my claim, compare the *Lancet* collection with books on research methods and statistical ideas such as Modern Epidemiology [5)] and Statistical Inference [6]. These are just two drops of teaching material in an ocean of textbooks and articles (that ocean is not captured in one black box called 'statistics'). There is also a sad secret to share: after mastering many of the topics, you would be able to voice an opinion on right and wrong in research *methods*, but you will still not know about truth or falsehood of research *results*. To find out why, you might wish to explore the interface between philosophy of science and statistics. There, you will find yourself in the midst of a battlefield among Bayesians (knowing is believing [7] and pure Likelihoodists which of two rival assertions is more likely? [8]) and Popperians (conjectural knowledge has nothing to do with beliefs [9]).

The answer to the second question (Will that treatment meaningfully help Mrs Smith?) is nowhere to be found [10], neither in the *Lancet* collection of essays, nor in any other writing. To understand why it cannot be found requires, again, a deep dive into fascinating metaphysical and methodological questions: (1) Is causation deterministic or probabilistic? [11]; (2) What is a causal parameter? [12]; (3) How many causal parameters are there for a given treatment and a given effect? [13]; (4) What is the epistemological status of a point estimate? [13]; and (5) How do we combine a fallible estimate of a causal parameter (even from a mega-randomized trial!) with auxiliary hypotheses to derive a treatment decision? None of these questions is systematically addressed in any of the essays. I think that a thoughtful discussion of these topics would have benefited the uninformed reader much more than a shouting headline about 'the scandalous failure of biomedical science to cumulate evidence scientifically' (p. 40). To read such a preaching title is scandalous indeed, I might add.

A thorough review of the *Lancet* collection of essays would require 20 commentaries, one per essay, which is unrealistic for a book review. In the next section, I will comment on one recurring theme; a few other issues will be addressed briefly in other sections.

External validity, generalisability, applicability, transferability, relevance, representativeness and similar wishful thinking

To understand the relation between a randomized trial (or for that matter, an observational study) and the unknown treatment effect in any particular human being, we need to step back and think about the estimated effect from a study. What exactly do we estimate in a sample? Are we merely estimating the average of deterministic effects, as some are quick to assume [12]? Are we estimating the average of stochastic effects? Or maybe we are estimating a homogenous probabilistic effect [13]? All of the essays skip the most crucial theoretical step, and many authors quickly jump into sweeping conclusions or personal convictions.

In panel 1 of the Preface (p. x), you will find a nice collection of prevailing dogmas: from 'Subgroup analysis kills people' (i.e. effects are never heterogeneous, so don't look for subgroup effects) to 'Anyone who believes that anything can be suited to everyone is a great fool' (heterogeneity of effects always exists). Neither view delivers an interesting idea, for a simple, technical reason. The trivial algebra of effect modification, which is usually taught in introductory epidemiology courses, shows the dependence of the phenomenon on the *scale* on which we choose to measure effects [14] (pp. 169–170). Some effect modification is *almost always* present on some measurement scale of effects, so the relevant question is not whether effects are ever heterogeneous, but whether we can argue that one scale is inherently superior in causal inquiry: for instance, can we argue that the rate ratio is preferred to the rate difference or vice versa? No, this volume never reaches that level of deliberation, but we do find all kinds of authoritative statements in many of the essays.

Some of the views about heterogeneity of effects seem to be anchored in methodological preference for null hypothesis testing (that null hypothesis testing has very few defenders against ample critics [15] is, unfortunately, ignored). According to this view, causal knowledge is accumulated in two steps: first, we assume that the effect must be zero until shown otherwise; next, we assume that a non-zero effect must be uniform until shown otherwise. That simplistic view is implied, for example, in the following quote:

> I want to know whether there are reasons that could justify confidently dismissing – as irrelevant to me – estimates of effects derived from the best available research evidence on groups of people. (p. 38)

May I ask where these reasons come from and what method allows us to make them not only reasons but also 'reasons that could justify confidently'? What judge do we nominate to issue the verdict on both 'justify confidently' and 'the best available research evidence'? May I suggest that the reasons for adopting (or dismissing) the estimates for a specific patient are simply untested conjectures about the absence (or presence) of effect-measure modification? May I suggest that what is called 'the best available research evidence' is decided by self-nominated interpreters of published research? Next: what logic tells me to give more weight to the homogeneity theory than to any competing theory of heterogeneity of effect? Why is one theory a priori more truthful than the other? Is it the same questionable logic that tells me to accept the null hypothesis as true until finding evidence to the contrary (Fisher), or the logic of a non-evidentiary, decision rule (Neyman-Pearson) [16]? Sadly, every treatment decision may be wrong, because it always requires auxiliary hypotheses [17], only some of which are testable. And that is definitely true – not only confidently true.

Another approach stops the heterogeneity train at the group level, describing external validity as being 'relevant to a definable group of patients' (p. 61). First, what exactly does 'relevant' mean in the writing of several authors? Does it mean equally effective for each member of the group, or are we back in the foggy idea called 'good on average', which implies 'bad or useless for some'? Second, as soon as one probe for the meaning of a 'definable group', it becomes obvious that no clear definition may be offered.

The only criterion to define a group of patients who would all identically benefit from some treatment is their sharing of the same causal parameter. Therefore, to define a group of patients to whom the treatment would be relevant, we need to know that the treatment effect is homogenous in that group. Do you share my feeling that we are trapped in a circular definition?

When it comes to heterogeneity of effects, everyone seems to have a pet theory to share. Who wants to argue against the possible heterogeneity of effect by the setting of care (p. 83), by disease severity (p. 144), by co-morbidity (p. 145) or by a risk score (p. 293)? Age has always been high on the list, especially old age (pp. 97–110), and every student of epidemiology learns these days about gene-by-environment interaction, namely, effect modification by our genes (pp. 144, 151–68, 307–18). Unfortunately, nature did not share with us her secret list of effect modifiers, so we often pursue them according to either technological advancements (e.g. genotyping) or sociological preferences (age, gender and ethnicity). Why pretend that scientific wisdom can guide the search?

All of these modifiers have a place in the world of scientific conjectures, although each of them could be wrong for any specific causal question (and some measurement scale). Indeed, regarding effect modification by place of care one author tells us that ‘it is not self-evident that trials undertaken in hospital are inapplicable to primary care’ (pp. 92–93). Sure, I agree. Everything goes in the domain of causal hypotheses, both conjectures of homogeneity and conjectures of heterogeneity. More important, in the very same essay we also find an illuminating paragraph about effect modification by personal identity:

> Ideally then, patients should undergo . . . n-of-1 trials to individualise their therapy. This has led to a recent revision of the ‘hierarchy of evidence’ for interventions, placing the n-of-1 trial at the top of the levels of evidence. These trials provide the best possible evidence for an individual patient . . . (pp. 86–87)

I cannot agree more with every word. All of a sudden, the discussion of external validity – from a classical randomized trial to Mrs Smith – is gone, and the possible heterogeneity of effect by personal identity is stated loudly and clearly. We discover that only technical reasons, such as cost or the nature of the effect of interest, prevent us from substituting an n-of-1 trial for every other source of causal knowledge, including systematic reviews. On a second thought, I wish the editor would have turned this volume into an open exchange between the various authors. They could have argued with each other on external validity, generalisability, applicability, relevance and the like better than I can.

Another prevailing perspective of external validity associates the idea with meeting the eligibility criteria of a trial (pp. 134, 139): if the patient were eligible to participate in a successful trial, the results apply. If not, then ‘the degree to which the RCT evidence applies to these patients cannot be assessed directly, and the clinician should not necessarily assume that their patient will respond to a medication in the same way as trial subjects’ (p. 134). That convoluted argument is false on several counts: first, there is no a priori way to assess directly or indirectly whether the result of a trial would apply to a single patient. It is merely a guess. Second, we do not know that all recipients of the tested treatment have responded identically (and we know nothing about the response of participants who did not get the tested treatment). Third, that a patient met (or did not meet) the eligibility criteria teaches us nothing about the effect of the treatment in that patient: a piece of paper on which we type a list of criteria for enrolment does not generate causal knowledge about the list [18]. Fourth, there is no way to tell what will happen to a particular patient. I repeat: no matter how well a treatment was tested and no matter how much external evidence has accumulated, there is no method to tell whether a given treatment will benefit, harm, or do nothing to a given patient. Those, who think they can tell the future, should go and buy the winning lottery ticket – every week.

The essay on pages 139–150 promises something unusual in science. It promises to teach you when to expect clinically important differences in the response to treatment. In other words, how to discriminate between theories of effect modification that will be proven true and those that will not. Moving one level up to the main effect, why not tell us the method by which we can discriminate between treatments that will work and those that will not? If there is a method to guess heterogeneity of effect, there should be an even simpler method to guess a non-heterogeneous effect, too. Of course, there is none, because the method of science is still bold conjectures and their rigorous testing. Arguments about the expected heterogeneity are just as good or just as bad, as arguments about the expected main effect, but the algebra is different. As I wrote earlier, ‘some effect modification is almost always present on some measurement scale’, so we don’t need to expect it. It is there. The real questions about effect modification are not discussed in these essays, and they are not statistical at all – at least for those of us who have expelled null hypothesis testing (including the Bayesian version) from science. Questions about effect modification are rooted in models of causation and epistemology.

Some people are convinced that they have found the ‘right scale’ to measure effects. For example, an endorsed quote tells us that:

> All policy [including treatment] decisions should be based on absolute measures of risk; relative risk is strictly for researchers only. (p. 247)

Naïvely, perhaps, I have always assumed that biomedical science should inform medical practice, but apparently, there are two kinds of research: research for the pleasure of the researchers (shall we call it the research game?) and research for the benefit of all other human beings. Along this line of thought, maybe we should ban all ratio-based regression models (logistic, Cox, Poisson), and declare them uninformative, irrelevant, misleading and damaging to treatment decisions and all other policy decisions. The essay on pages 247–263 does not dare to say so explicitly, but I see no other interpretation of the text.

According to which model of causation differences are superior to ratios? Determinism? How do we know that the model is correct? What if the model is wrong? What are the consequences of this model for causal inference? What is the conceptual difference between the idea of deterministic interaction and the idea of indeterministic effect-modification? No need to bother about these fundamental questions that lie at the core of causal inquiry: all treatment decisions should be based on the risk difference. One problem was solved. What is the next one?

The play of chance

There is nothing I dislike more than the word ‘chance’ in discussions of cause and effect. The reason? It is a helpful statistical device to hide whatever one wishes – a wild card for all types of ignorance. We all accept chance as part of causal reality (sometimes thinking probabilistically), yet we also often switch to determinism without worrying too much about incoherence or inconsistency.

What sounds more truthful than 'many patients are needed to distinguish true treatment benefits . . . moderate in size, from chance effects' (p. 183)? Not so. May I ask what statistical theory establishes any estimate as true or false? What statistical theory tells us that a point estimate from a large trial is the true effect whereas a point estimate from a small trial is a 'chance effect'? Is this another example of an erroneous interpretation of a small *P*-value as evidence for the truth of the estimate [19]? Or maybe an erroneous interpretation of a confidence interval [20]? Are we forgetting again an embarrassing statistical distinction between a fallible estimate and the unbiasedness of an estimator, which is the property of any valid randomized trial, including a tiny one [21]? Reading more into the essay, I think I found the answer: 'evidence for differential effects should be assessed by a statistical test of interaction' (p. 187). Someone is convinced that causal knowledge must follow the *P*-value ritual [22].

Confusing prediction with causation

In my years of studying and teaching epidemiology and biostatistics, I have frequently witnessed a common confusion between two distinct ideas: estimating the effect on the outcome and predicting the outcome. The confusion arises from failure to understand the dual role, and sometimes conflicting role, of regression analysis. On one hand, we may construct a regression model with a causal analysis in mind: we want to estimate the *effect* of a variable on the outcome (using a regression coefficient). On the other hand, we may construct a regression model with *prediction* in mind. For example, wishing to know the probability that Mrs Smith would die in the next year, we search for a set of easily measured regressors (independent variables) that would *collectively* predict death 'well enough'. Then, we plug Mrs Smith's values of the regressors and compute a probability.

Two key points should be borne in mind about the second goal (prediction): first, although causes of the outcome often find their way into the regression model, a good prediction model does not need to include only causes. Variables that do not affect the outcome could help to predict it, even better than measurements of the true causes. If teenagers, for instance, tend to lie about their smoking status, yet tell the truth about possessing a cigarette lighter, we would predict their probability of developing lung disease much better from the non-causal variable 'possessing a cigarette lighter' than from the measured variable 'smoking status'. Second, for reasons that originate in the theory of directed acyclic graphs [23,24], an excellent prediction model might do a poor job in estimating the effect of some, or all, of the causal variables. Unfortunately, so many of us, including experienced statisticians, seem to confuse the asymmetrical relation between causal inquiry and statistical prediction: to estimate a causal parameter, we often seek help from a prediction model (regression), but not every prediction model – no matter how good it may be in predicting the outcome – delivers unbiased estimators of causal parameters. Therefore, to talk about 'predicting the effect' is an erroneous hybrid of different ideas.

The *Lancet* collection contains two essays on prediction: 'Use of risk models to predict the likely effects of treatment in individuals' (pp. 195–211) and 'Evaluating the performance of prognostic models' (pp. 213–229). I will let you try to reconcile these essays with what I wrote above. It may not always be easy.

Are they teaching you epidemiology?

Students who take a course in introductory epidemiology are taught that effects are not measured by computing the 'absolute risk'. They are taught that effects must be defined on pairs of causal contrasts, and that effects may be estimated by subtraction or by division. For instance, the rate difference and the probability difference are difference measures of effect; whereas, the rate ratio and the probability ratio are ratio measures of effect. Then, astute students often ask which scale and which measure is preferred. I tell them that the issue is too complicated to be explained in an introductory course [25,26]. Later in the course, when we discuss effect modification, I show them how the algebra can generate different inference depending on whether we compute a difference or a ratio, and they ask again: is one scale preferred to the other? If there is no clear rationale for choosing one scale, what's the point of talking about effect modification, anyway? I repeat the same answer.

The story I have just told you is a typical example of the complexity of research methods. At every turn – whether it is a measure of effect, or the meaning of confounding, or the logic of matching controls to cases – we find deep complexity behind apparently simple ideas. Ironically, it is easy to hide the complexity by a careful choice of words. For example, rather than discussing the difficult methodological choice between difference and ratio measures of effect, you can discuss 'absolute risk reduction' and 'relative risk', which are powerful synonyms. The words will do the work for you because 'relative' surely sounds inferior to 'absolute' (pp. 141–142, 247–263). Or another example: with trivial math, you can convert the probability difference for death into 'the number needed to treat to save one life', and have the rhetorical power on your side [27].

Here and there, the *Lancet* collection includes rudimentary attempts to teach the complexity of epidemiological ideas (e.g. pp. 16–29). Sorry, but I don't believe in crash courses. There is no short substitute for rigorous, time-consuming training in observational research, just as there is no short substitute for rigorous training in medicine. In my view, it is pretentious and irresponsible to teach epidemiological ideas and methods 'on the side', as part of a collection on randomized trials and medical practice. Naïve readers might get the false impression that they know more than what they can actually know.

Anything positive?

I did find a few pearls in the *Lancet* collection, one of which is noteworthy. When the title of an essay reads 'applying results to treatment decisions in complex clinical situations' (p. 111), you know what is waiting ahead. No fan of the rhetoric of *evidence-based medicine*, *best evidence*, *best practice*, *managed-care*, *systematic reviews* and *practice guidelines* will choose such a title. So please join me in reading the words of a doctor who had to decide whether to prescribe warfarin to one of his patients:

> Using these [research] data, my own anecdotal experience, and information about the pathology and pathophysiology of brain ischaemia in general and in Asians, I decided that warfarin anticoagulation was probably the most effective medical treatment . . . I decided, despite the relative contraindications, to try anticoagulant treatment . . . Did I make the correct decision? Time will tell how she does . . . (p. 117)

Did he use any algorithm, prescribed rules or dogmatic thinking to decide on 'external validity'? No. Would you want him to be your doctor? I would. No false claims to know the truth about cause-and-effect and no statements about 'the lethal consequences of failing to make full use of all relevant evidence about the effects of medical treatments: the importance of systematic reviews' (p. 37). For sure, most medical decisions are made in a complex situation, but sometimes the complexity variables are unknown or are unmeasured. It may be easy to know that the patient does not take drugs as prescribed; it is difficult to know how quickly a drug will be metabolized in the patient's liver.

There is much to be learned from comparing the story of that doctor with a rule of treatment issued in another essay – apparently written by a statistician:

> When we do not have evidence about treatment effects in specific subgroups of patients, these decisions have to be based on evidence about overall effectiveness. We should only make different decisions for specific patient groups when strong evidence supporting this becomes available. (p. 191)

May I ask about the theoretical basis for that authoritative prescription? Has he never seen a human being who was harmed when that rule was followed? Upon further inspection of the text, I think I recognize the differences between the mind of the doctor I quoted earlier and the mind of the statistician: the first one was making a treatment decision for a single patient; the second sees only 'patient groups'. The first one was worried about the fallibility of scientific knowledge and possible heterogeneity by personal identity; the second denies any heterogeneity until 'proven' otherwise (by a small *P*-value, maybe?). The first one was not sure about his treatment decision for a single patient; the second has no doubt about how all patients should be treated. Which school of thought do you prefer in medical practice?

What will you take home after reading the book?

Over the past few years, I have been trying to write my own fallible thoughts on the study of cause-and-effect. From this book-in-progress, I chose to share the concluding part of the Preface – because it offers a strikingly different perspective from the premises and promises of the *Lancet* collection. The following is taken from a section titled 'What will you take home?'

> You will take home, I hope, the following four messages: First, like all scientific knowledge, knowledge of causes remains conjectural forever, no matter how strongly we believe otherwise. Second, learning about causes should be equated with estimating the magnitude of their effects, not with declaring them causes. Third, every estimate of a causal effect embeds untestable assumptions; hence the first message. Four, there are better methods and poorer methods for causal inquiry: there are methods that rest on good reasoning, and there are methods that don't. Keep in mind, however, that good reasoning does not guarantee a hit on the truth, and that's true for every branch of science. We, scientists, constantly seek the Truth but never know whether we have reached it. If not our job, at least our profession is secured.

Unless I missed it, I have not read anything close to these ideas in the *Lancet* collection. In fact, the editor, just like the *Lancet* editors a decade ago [28], is hopeful that some day we will solve 'the main challenge still facing evidence-based medicine – how best to inform the treatment of individuals' (p. xi). I am sure he does not mean how to force a doctor to follow the latest verdict that was issued by the latest author of the latest systematic review.

Buy the book!

Publishers fear critics because research has shown that a harsh critique adversely affects the sales of a book. In the spirit of the topic at hand, I would like to make those research results *externally invalid*, *non-generalisable*, *inapplicable* and *irrelevant* to my own critique. Therefore, I urge you to purchase and read the *Lancet* collection of essays – if only to tell me whether you disliked the book as much as I did.

References

1. The *Lancet* (2007) Treating Individuals: From Randomised Trials to Personalised Medicine. Edinburgh: Elsevier.
2. Miles, A., Loughlin, M. & Polychronis, A. (2007) Medicine and evidence: knowledge and action in clinical practice. *Journal of Evaluation in Clinical Practice*, 13, 481–503.
3. Shahar, E. (1998) Evidence-based medicine: a new paradigm or the Emperor's new clothes? *Journal of Evaluation in Clinical Practice*, 4, 277–282.
4. Shahar, E. (2003) Generalizability: beyond plausibility and handwaving. *Journal of Evaluation in Clinical Practice*, 9, 151–159.
5. Rothman, K. J. & Greenland, S. (1998) Modern Epidemiology. Philadelphia: Lippincott-Raven.
6. Oakes, M. (1990) Statistical Inference. Chestnut Hill: Epidemiology Resources Inc.
7. Howson, C. & Urbach, P. (1993) Scientific Reasoning: The Bayesian Approach. Chicago and La Salle: Open Court Publishing Company.
8. Royall, R. M. (1997) Statistical Evidence: A Likelihood Paradigm. Boca Raton: Chapman & Hall/CRC.
9. Miller, D. (1994) Critical Rationalism: A Restatement and Defence. Chicago and La Salle: Open Court Publishing Company.
10. Shahar, E. (1997) Translating clinical trials into practice. *Lancet*, 349, 654–655.
11. Hallett, M. B. (1997) Is 'life' based on clockwork biology or quantum uncertainty? *Perspectives in Biology and Medicine*, 41, 101–107.
12. Maldonado, G. & Greenland, S. (2002) Estimating causal effects. *International Journal of Epidemiology*, 31, 422–429.
13. Shahar, E. (2007) Estimating causal parameters without target populations. *Journal of Evaluation in Clinical Practice*, 13, 814–816.
14. Rothman, K. J. (2002) Epidemiology: An Introduction. New York: Oxford University Press.
15. Parkhurst, D. F. (1997) Commentaries on significance testing. Available at: http://www.indiana.edu/~stigtsts/index.html#contents (last accessed 23 January 2008).
16. Hubbard, R. & Bayarri, M. J. (2003) Confusion over measures of evidence (*p's*) versus errors (α*'s*) in classical statistical testing. *The American Statistician*, 57, 171–178.
17. Shahar, E. (2003) On morality and logic in medical practice: commentary on 'A critical appraisal of evidence-based medicine: some ethical

considerations' (Gupta 2003; Journal of Evaluation in Clinical Practice, 9, 111–121). *Journal of Evaluation in Clinical Practice*, 9, 133–135.

18. Senn, S. J. (1991) Falsificationism and clinical trials. *Statistics in Medicine*, 10, 1679–1692.
19. Shahar, E. (2007) Commentary: interpreting the interpretation. *Journal of Evaluation in Clinical Practice*, 13, 693–694.
20. Poole, C. (1987) Confidence intervals exclude nothing. *American Journal of Public Health*, 77, 492–493.
21. Greenland, S. (1990) Randomization, statistics, and causal inference. *Epidemiology*, 1, 421–429.
22. Cohen, J. (1994) The earth is roun*d (P < 0.05). American Psychologist*, 49, 997–1003.
23. Greenland, S., Pearl, J. & Robins, J. M. (1999) Causal diagrams for epidemiologic research. *Epidemiology*, 10, 37–48.
24. Hernan, M. A., Hernandez-Diaz, S. & Robins, J. M. (2004) A structural approach to selection bias. *Epidemiology*, 15, 615–625.
25. Greenland, S. (1987) Interpretation and choice of effect measures in epidemiologic analyses. *American Journal of Epidemiology*, 125, 761–768.
26. Walter, S. D. (2000) Choice of effect measure for epidemiological data. *Journal of Clinical Epidemiology*, 53, 931–939.
27. Cook, R. J. & Sackett, D. L. (1995) The number needed to treat: a clinically useful measure of treatment effect. *British Medical Journal*, 310, 452–454.
28. Editorial (1997) And now all this. *Lancet*, 349, 1.