Clinics in Dermatology (2013) **31**, 806–810

ELSEVIER

Clinics in Dermatology

COMMENT AND CONTROVERSY
Edited by Stephen P. Stone, MD

Why we should let "evidence-based medicine" rest in peace

Jonathan Rees, FMedSci*

University of Edinburgh, EH3 9HA Edinburgh, Scotland

Abstract Evidence-based medicine is a redundant term to the extent that doctors have always claimed they practiced medicine on the basis of evidence. They have, however, disagreed about what exactly constitutes legitimate evidence and how to synthesize the totality of evidence in a way that supports clinical action. Despite claims to the contrary, little progress has been made in solving this hard problem in any sort of formal way.

The reification of randomized clinical trials (RCTs) and the tight linkage of such evidence to the development of clinical guidelines have led to error. In part, this relates to statistical and funding issues, but it also reflects the fact that the clinical events that comprise RCTs are not isomorphic with most clinical practice. Two possible and partial solutions are proposed: (1) to test empirically in new patient populations whether guidelines have the desired effects and (2) to accept that a distributed ecosystem of opinion rather than a hierarchical or consensus model of truth might better underwrite good clinical practice.

Introduction

Thirty years ago John Hampton, a British cardiologist, wrote an editorial, in the *British Medical Journal* with the stark title, "The end of clinical freedom.[1]" It began, "Clinical freedom is dead, and no one need regret its passing." Hampton believed that support for clinical freedom was all too often a cloak for ignorance, and that if for no other reason, rising health care costs made ever more pressing the need for randomized clinical trials to underpin—perhaps even dictate—clinical practice.

Almost thirty years later, when the *International Journal of Epidemiology* reprinted and published commentaries on his now-famous editorial, Hampton saw things differently.[2] The new contribution was now titled, "The need for clinical freedom," and in it, he reviewed what had happened to his earlier vision of how large RCTs were to improve medical practice. He ended this latter essay as follows:

> So we seem to have the perfect storm, where a meeting of evidence-based (which we ought to call opinion-based) proscriptive guidelines, mechanistic doctors and financial control have come together to contribute to the demise of the responsibility that doctors used to have for individual patients. We need to change medical culture in such a way that doctors can use their opinions about published evidence to select the best treatment for each individual patient. We need a return to clinical freedom.

The issues Hampton grappled with are as foundational for medicine as any: what knowledge underpins clinical expertise and clinical practice; how do we acquire this knowledge; and how is this knowledge validated within the

* Corresponding author. Tel.: +44 0 1315362041.
E-mail address: reestheskin@me.com (J. Rees).

0738-081X/$ – see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.clindermatol.2013.06.004

framework of a profession? How can we—and others—have trust in what we claim to know?

From Paracelsus to evidence-based medicine

Ian Hacking, a philosopher of both science and probability, describes the evidence that underpinned Paracelsus's assertion that mercury was an effective treatment for syphilis.[3] Hacking prefaces the argument as follows:

> Syphilis is signed by the market place where it is caught; the planet Mercury has signed the market place; the metal mercury, which bears the same name, is therefore the cure for syphilis.

Well, of course, this makes absolutely no sense to the modern mind. We simply do not accept the validity of the concept of entities being "signed" as a legitimate form of evidence. So, when we read one of the most widely quoted definitions of evidence-based medicine (EBM)[4,5] as "… the conscientious, explicit, and judicious use of current best evidence in making decisions about the care of individual patients," we are forewarned. The key issue is not whether we use evidence or not, but exactly what constitutes legitimate or "current best" evidence. EBM is a redundant term, because doctors have always justified their behavior on the basis of evidence. What they have disagreed about is what constitutes legitimate evidence and, in particular, how doctors should make trade-offs between different sorts of evidence.

Evidence-based medicine: Demarcating acceptable evidence

EBM, therefore, needed to demarcate acceptable from unacceptable forms of evidence. The solution was to invent a hierarchy of quality of evidence and to use numbers to describe categories of methods, with evidence from randomized controlled trials (RCT) being held as the gold standard.[5] This widely used metaphor—that of the gold standard—needs unpacking[6].

If you peg the dollar or the Euro to a gold standard, you promise that you are willing (in theory at least) to exchange a defined amount of gold for your paper currency. Unless you suddenly acquire large amounts of gold (at low or zero cost), your ability to print money without devaluing your currency is curtailed; however, for this metaphor to work in the current context, we have to either equate the RCT with truth (as a matter of definition) or just say that we want to be able to exchange various forms of non-RCT clinical evidence with the RCT gold standard.

In any formal sense (as in the sense that mathematics is formal), nobody knows how to create an explicit procedure for converting different types of evidence. All you can do is exercise judgment and opinion: in medicine we call this clinical judgement.[7] The approach taken by most of the EBM faithful, especially those who do not practice medicine, has been to try to ignore all forms of evidence that are not RCT.[5,8] This is operationally attractive but comes at the expense of excluding most evidence, a position that seems a little curious for those boasting about their respect for empirical evidence. Historically, pointing a loaded revolver at one's head and pulling the trigger is not a risk-free activity. In the absence of a RCT, however, we are forced to overlook history's evidential value and merely report that our literature searches failed to reveal any high-quality evidence.

The focus on RCTs leads to other problems. Much as we lack a metric to exchange non-RCT evidence for RCT-evidence, we also have to assume that not only are RCT infallible, but that they are also isomorphic with the everyday clinical encounter.[6] Trials are not viewed as guides or metaphors that we can use to inform everyday practice, but in a curious reversal, everyday practice is deemed to be equivalent and subservient to the trial conditions. What is true in a RCT must be true in clinical practice, because good "clinical practice" has been operationally defined as that which takes place in a RCT. This is, of course, another sleight of hand, but one that at least initially made life very difficult for those who suspected that the emperors of EBM were, if not naked, scantily dressed.[9]

As trials multiplied, it became clear that apparently well-designed RCTs on the same topic often came up with different results. In one sense, this should not have surprised anybody. In many clinical trials, the effect of an intervention is remarkably small, and therefore statistically unstable. A sensible solution—if the studies really did address the same topic in a way that made sense to combine them—was to perform a meta-analysis. Again, although the evidential net was being cast wider, non-RCT evidence was usually to be excluded from most systematic reviews (although meta-analyses of other study designs are possible too). If the individual trials could not be relied on, could the meta-analysis take on the role of a now newly minted gold standard?

The answer to this question came out of left field. The Dutch clinical epidemiologist, Jan Vanderbrouke, drew attention to an apparently state-of-the-art meta-analysis of homeopathy, which suggested that homeopathy worked better than placebo.[10] Vandenbroucke, with considerable imagination, realized that while this meta-analysis told us relatively little about homeopathy, it told us a great deal about RCTs and meta-analysis itself. Something important was wrong.

Vandenbroucke argued as follows.[10] In the meta-analysis, homeopathy was deemed an active treatment and was compared with a control. He pointed out that because homeopathic agents are so dilute, they contain no active agent but are in reality just another control. This meant that the meta-analysis now showed—in an apparently statistically

robust way—that two different inactive controls differed significantly. All within the framework of best practice!

The take-away message was simple. Trials and meta-analyses are much noisier forms of evidence than many had imagined.[10] None of this should have been too surprising. As has been said on many occasions, "significance tests are for situations where we do not understand, in any theoretical sense, what is happening."[11] By contrast, physicists use statistics in their theories, but rarely use them to test whether their theories are true or false.[12] Fisher's invention of much of the modern statistical armamentarium was to allow him to quickly triage lots of data that had been collected with little experimental design. Small *p*-values might mean there were things of interest in the results, but Fisher realized *p*-values were not markers of truth (otherwise, we would have to say he had invented a truth machine, something as improbable as a perpetual motion machine).[12]

RCT versus reliable knowledge

The easiest way to appreciate the problems that the fixation with RCT has led to is to imagine how or why we could be confident of an effect without a RCT. Rather than use the earlier Russian roulette example, think of a clinical intervention such as excision of basal cell carcinomas. How many patients would you need to see and treat before you were convinced the treatment worked? Surely, a handful, at most! The efficacy is such that it is robust to alterations in lots of clinical factors. Yes, surgical expertise might influence the results a little. Yes, occasional large and aggressive tumors might not be treatable. Yes, new primaries close to the first tumor might be mistaken for failures of the primary excision, and so forth. What is striking, however, is that the magnitude of effect is such that everyday practice (literally) provides confirmatory evidence of what we believe, and that however you measure the clinical endpoints, few would doubt efficacy. Here is a key point: our beliefs are reliable and robust due to what we see every day, despite the variation in patient mix or subjective assessment methods that we all use in normal day-to-day practice. There are plenty of other examples too: systemic isotretinoin for acne, dapsone for dermatitis herpetiformis, anthralin or cyclosporine for psoriasis, and so forth.

RCT and the allure of small effects

All of the examples quoted in the previous section were of treatments that have large effect sizes (ie, treatments that work well). Most patients who receive the active intervention get some benefit, and everyday practice provides a useful guarantor of what we might have read in the journals.

For many interventions, this is not the case. Effect sizes are small, and relatively few patients benefit. When effect sizes are larger and therefore easily detectable when compared with placebo, in head-to-head comparisons with other active agents, the differences will be more modest. Examples in dermatology would include the size of surgical margins for melanoma or whether some basal cell carcinomas should be treated with curettage rather than excision. Here, while the likely effect sizes of either intervention are large, the power to detect differences between different active interventions in everyday practice is limited.

It is, of course, in these situations that RCTs are both attractive and powerful. If we could conduct large-scale RCTs using patients from our practices over the long term and using the sorts of measures that we believe are important to our patients, this essay would end here. Note all the implied caveats in the previous sentence. The problem is that we seldom can conduct such RCTs.[8]

Discovering treatments that work very well is much harder than discovering treatments that work less well, and the allure of the large-scale RCT was that we could measure effect sizes for treatments that do not work very well. It is this more than anything else that has led to so many of our current problems. The difficulty with small effects is that they are difficult to detect and lack robustness (meaning that if we repeat the study again and again, we may see different results). We could combine the results of studies in a meta-analysis, but evidence shows that even when this is done, subsequent larger studies often show different results.[2]

As we scale up the numbers needed for a large study, the cost of studies become ever greater, and concerns (including reputational or financial) about getting a significant result ever greater. The entry criteria become more and more limited, and the geographical study area larger and larger. The result is that the average patient we see in practice resembles less and less the patients enrolled in trials. Does this matter? Yes, it seems to.[13] (What of course we would like to do is enroll patients with reference to a geographically defined population as has been suggested, but this would likely demand longer and more expensive study periods).

Most large studies rely on assessment methods and rating scales that do not necessarily reflect either our values or those of our patients.[8,9] The scales are a tool to reduce statistical variation but at the cost of clinical veracity. Is complete clearing of psoriasis, reduction in psoriasis on visible areas, or percentage reductions in the psoriasis area severity index (PASI) what your patient wants? Side effects of many drugs may vary from easily detectable and reversible by stopping, through catastrophic and (largely) irreversible. RCTs do a poor job of picking up rare side effects, nor are they designed to do so. It is salutary to remember that thalidomide was the first drug to get a U.S. Food and Drug Administration (FDA) license based on RCTs conducted prior to licensing.[8]

Then, there is an interaction between efficacy and safety. In a trial of different treatments for pemphigoid, for example, do we use blisters or death as the primary outcome? The two measures may be inversely related, so must we power all studies to detect the rarer outcome, death? How do our

patients make this trade-off, and do they make it the same way? Finally, although we would prefer drugs that work very well in all patients, some drugs that work less well on average appear to work very well for some patients.[9] A topical vitamin D agonist may only work so-so on average, but the average might only be average, because some people respond very well, and others not so. In practice, in this example, you might try it and see, as you get to learn from the same patient over time. In most trials (because trials with *n* of 1 are vanishingly rare), all you see is the average effect.

The pressure for positive results

The undoubted experimental power of designs that incorporate randomization and control groups lulled us into forgetting that the most important determinant of confidence in a drug's action is effect size: does it work well in most patients, and is this finding robust enough to see in everyday practice. The cost and complication of testing for small effects has resulted in an ecosystem that has become ill suited to what Hampton in his first paper had wished.[8] We know that many studies remain unpublished, that national guidelines are based on inadequate access to existing trials, and that many trials use outcomes of only limited relevance to patients (or doctors). If we are to believe John Ionnadis, most published studies are wrong due to an interaction between commonly used statistical methodologies (the lack of strong priors) and selection effects on what is submitted and what is accepted for publication.[14]

As David Healy has pointed out, there is a considerable irony here.[8] Clinical trials were introduced in part to challenge companies to produce evidence of efficacy (as opposed to only demonstrating safety), but the reification of RCT means that if you can design and determine what RCTs are undertaken and published, then you control the evidential landscape base. Because it is the makers of a drug that fund most clinical trials, they partly determine the outcomes of apparently independent guideline committees. Anything outside the accepted evidential norms is mere "anecdote" and can be safely ignored: lies, damned lies, and clinical impressions.

Two computing analogies that might shed light on how to improve care

Getting out of this mess first means understanding how we got into it. Assessing efficacy of agents in routine clinical practice is hard for agents that work less well, especially if the natural history is variable. If the effect size is small, it is nigh on impossible. Similarly, if side effects are rare, we are unlikely to assess them well in RCTs (nor in everyday clinical practice). We will need to amalgamate clinical experience using formal recording systems, but we will still have to argue about possible confounding. The attraction of RCTs is that they create statistical power and allow us to worry less about biased assignment of patients to alternative treatments.

Trials have their downsides too, something that one of the fathers of the modern RCT, Austin Bradford Hill, said almost fifty years ago: "Any belief that the controlled trial is the only way [we could study therapeutic efficacy] would mean not that the pendulum had swung too far but that it had come right off its hook."[15] One thing we can do is to stop reifying RCT evidence over other types of clinical knowledge. That is, we need to break the rigid links between RCT results and guideline development on the one hand, and clinical practice on the other.[13]

Clinical guidelines are essentially a set of semiformalized instructions in a very high-level language that attempt to map inputs onto outputs—to map patient clinical information with therapy or clinical action. In this sense, they are analogous to computer code. Rather than the current emphasis on process,[16] the crucial measure we should be concerned with is whether the code works—that is, does it produce the desired outputs for a range of legitimate inputs. Computer coders know that most attempts to make explicit any series of instructions result in errors. How you write the code, what language you use, and so forth are less important than its behavior in mapping input to output.

A textbook is a guideline, much as are the range of pithy heuristics used in medicine (eg, "bleeders come first"). In deciding which to choose, we need to measure performance in the real world; the default is that guidelines will always contain bugs that may result in worse clinical outcomes in some situations, and that the performance of different guidelines will vary independently of defined inputs. To discover such bugs, in addition to asking what colleagues think, one must run the code in the real world of the clinic; note that the former is not a substitute for the latter. Differences between guidelines and practice may therefore reflect a range of problems: individual clinicians may be at fault in ignoring guidelines; the data or opinions on which guidelines are based may be in error; or the attempt to formalise knowledge may be bug-ridden with unintended outputs. Empirical scrutiny of outcomes—not badging of process—is what is needed.

Another computing metaphor might be useful at this stage, but it points us in a very different direction. Based on the paradigm that pharma companies have to follow to obtain drug registration, we have assumed that guides to clinical practice have to be hierarchical and bureaucratically "quality assured." This is most obvious in countries such as the United Kingdom (UK), where the state wishes to be the sole arbiter of how people are treated (in part because much health care is tax-payer funded but also because the state likes to assert control of health). The World Wide Web offers us another model of expertise, one in which the idea of a single central authority assuring the truth or falsity of statements has been replaced by a community—or cacophony, depending on one's viewpoint—of voices.

Here expertise is distributed, and the measures of truth are perhaps much more nuanced and fluid, subject to change as data and clinical experience changes. Curiously, it is this latter model, albeit using earlier methods of communication, that was the basis for the growth of scientific ideas and our interpretation of evidence about the world. It might be worth revisiting.

References

1. Hampton JR. The end of clinical freedom. *Br Med J.* 1983;287:1237-1238.
2. Hampton J. Commentary: The need for clinical freedom. *Int J Epidemiol.* 2011;40:849-852.
3. Hacking I. *The emergence of probability: A philosophical study of early ideas about probability, induction and statistical inference*. Cambridge: Cambridge University Press; 1984.
4. Sackett DL, Rosenberg WM, Gray JA, Haynes RB, Richardson WS. Evidence based medicine: What it is and what it isn't. *BMJ.* 1996;312:71.
5. Charlton BG. The zombie science of evidence-based medicine: A personal retrospective. A commentary on Djulbegovic, B, Guyatt, GH, & Ashcroft, RE (2009). *Cancer Control*. 2009;16:158-168 and *J Eval Clin Pract*. 2009;15:930-934.
6. Rees JL. The nature of clinical evidence: Currencies floating rather than gold standards. *J Invest Dermatol.* 2007;127:499-500.
7. Feinstein AR. Clinical judgement. Baltimore: Williams & Wilkins Company; 1967.
8. Healy D. *Pharmageddon*. Berkeley: University of California Press; 2012.
9. Rees JL. Trials, evidence, and the management of patients with psoriasis. *J Am Acad Dermatol.* 2003;48:135-143.
10. Vandenbroucke JP. Homoeopathy trials: Going nowhere. *Lancet.* 1997;350:824.
11. Hacking I. *An introduction to probability and inductive logic*. Cambridge: Cambridge University Press; 2001.
12. Gigerenzer G, Swijtink Z, Porter T, Daston L, Beatty J, Kruger L. *The empire of chance: How probability changed science and everyday life*. Cambride: Cambridge University Press; 1990.
13. Hampton JR. Guidelines—for the obedience of fools and the guidance of wise men? *Clin Med.* 2003;3:279-284.
14. Ioannidis JPA. Why most published research findings are false. *PLoS Med.* 2005;2:e124.
15. Hill AB. Reflections on the controlled trial. *Ann Rheum Dis.* 1966;25: 107-113.
16. Kahn R, Gale EA. Gridlocked guidelines for diabetes. *Lancet.* 2010;375:2203-2204.