The evolution of evidence hierarchies: what can Bradford Hill's 'guidelines for causation' contribute?

Jeremy Howick[1] • Paul Glasziou[1] • Jeffrey K Aronson[2]

[1] Centre for Evidence-Based Medicine, Rosemary Rue Building, Old Road Campus, University of Oxford, Oxford OX3 7LF

[2] Department of Primary Health Care, University of Oxford, Oxford

Correspondence to: Jeremy Howick. E-mail: Jeremy.howick@dphpc.ox.ac.uk

DECLARATIONS

Competing interests
None declared

Funding
None

Ethical approval
Not applicable

Guarantor
JH

Contributorship
This paper was a truly collaborative effort that resulted from a series of meetings attended by all three authors. JH produced the initial draft and was in charge of revising subsequent drafts. PG provided insights about the Mother's Kiss example, and was also instrumental in conceptualizing the last diagram. JKA was instrumental for the adverse drug reaction example and also in coming up with ideas for revising the 'dose-response'

'A main cause of philosophical disease – a one-sided diet: one nourishes one's thinking with only one kind of example.' Ludwig Wittgenstein

Introduction: when non-RCT evidence is sufficient to conclude that the intervention caused the outcome

High quality randomized controlled trials (RCTs) (concealed allocation, relevant groups blinded and sufficiently powered, etc.) will usually provide sufficient evidence to establish that a particular treatment caused an outcome. Yet sufficiently well-conducted RCTs are rare.[1] Trials can be under-powered,[2] or unsuccessfully blinded,[3,4] and often suffer from many undetected biases. The results of most RCTs are therefore often insufficient to establish causation. At the same time, RCTs are often not required to establish causation.[5] Treatments including the Heimlich manoeuvre, cardiac defibrillation and parachutes to prevent death[6] have never been tested in RCTs, yet their effectiveness is surely strongly supported by evidence.

Evidence-grading systems that place randomized trials at the top of a hierarchy[7–13] will deliver misleading conclusions in cases where RCTs are insufficient or unnecessary. According to these hierarchies, trails of homeopathy – often generating positive results and generally of higher quality than RCTs of conventional treatments[14] – will be considered to provide strong evidence, whereas the evidence base for the Heimlich manoeuvre to unblock airways and parachutes to prevent death will be judged as less strongly supported by evidence.

Sir Austin Bradford Hill, in a widely-cited 'pre-EBM' system for appraising evidence, suggested that several relevant factors must be considered before concluding causation. We investigated and revised the Bradford Hill 'guidelines for causation', in order to refine our intuitions about whether to believe that intervention is effective. Our intention is not to debunk previous attempts to grade evidence, but rather to contribute to their natural evolution and development.

Revising Bradford Hill's guidelines

We believe that Bradford Hill's guidelines form a useful tool as they stand. Nevertheless, they can be modified in ways that make them easier to use. For instance, some of the guidelines, such as 'specificity' can safely be omitted while others, such as 'experiment' and 'strength' can be combined; still others, such as 'biological plausibility' can benefit from a more detailed analysis. Moreover, the guidelines have an inherent structure that is unclear in the original exposition. We propose that the guidelines be organized into the following three categories:

(1) *Direct evidence* from studies (randomized or non-randomized) that a probabilistic association between intervention and outcome is causal and not spurious;
(2) *Mechanistic evidence* for the alleged causal process that connects the intervention and the outcome;
(3) *Parallel evidence* that supports the causal hypothesis suggested in a study, with related studies that have similar results.

A previous attempt to impose a structure on the guidelines[15] may have oversimplified, claiming, for example, that 'analogy' (our 'similarity') is a 'mechanistic' consideration (which, as shall become clear below, is a category error).

guideline. He was also responsible for the suggestion to combine and omit some of the guidelines

Acknowledgements

We are grateful to Nancy Cartwright for providing useful insights during conversations with the authors. Members of the GRADE working group, especially Roman Jaeschke and Joseph Watine, provided useful feedback. Murray Enking read an earlier draft and suggested the example of folic acid to prevent neural tube defects

We use the term 'guidelines' over the more common 'criteria'[16–21] because Bradford Hill did not regard any of the guidelines as necessary or sufficient for establishing causation[11]: '… none of these viewpoints can bring indisptuable evidence for or against a cause-and-effect hypothesis and equally none can be required as a *sine qua non*'.[22] To cite his example, 'It will be helpful if the causation we suspect is *biologically plausible*, though this is a feature we cannot demand. What is biologically plausible depends on the biological knowledge of the day.'[22] Bradford Hill gave similar warnings about all the other guidelines (except, as we shall see, 'temporality'). Rather than 'criteria', they are best viewed as factors to be considered when assessing whether there is evidence for causation, or 'guidelines' for short.

Aware of detailed descriptions of the original guidelines,[15,23,24] we shall limit ourselves to describing our re-structured and revised version (Table 1). We shall then apply the Revised Bradford Hill Guidelines to real examples of likely causation despite lack of support from RCTs.

Direct evidence

The first three of the revised guidelines help assess whether 'direct' evidence of a probabilistic association between two factors is causal rather than spurious.

Table 1
Bradford Hill's original guidelines and proposed revisions

Type of evidence	*Revised, structured guidelines*	*Hill's original guidelines*
Direct	Size of effect not attributable to plausible confounding	Experiment
	Appropriate temporal and/or spatial proximity (cause precedes effect and effect occurs after a plausible interval; cause occurs at the same site as the intervention)	Strength
	Dose-responsiveness and reversibility	Temporality
Mechanistic	Evidence for a mechanism of action (biological, chemical, mechanical)	Biological gradient Biological plausibility
Parallel	Coherence	Coherence
	Replicability	Consistency
	Similarity	Analogy

Size of effect not attributable to plausible confounding

Plausible confounders are factors which are not directly related to the experimental intervention, are unequally distributed between treatment and control groups, and are likely to determine the outcome. For instance, we might observe that depressed people who exercise recover more quickly. Is the association between exercise and more expedient recovery from depressive symptoms causal? We cannot answer this question without ruling out potential confounders. Those who take regular exercise might also (on average) get more sun, eat healthier foods or they might simply believe more strongly that their depression will go away. These other factors, rather than exercise, might cause their speedier recovery.

Different ailments and studies are at risk from different confounders, so the judgement of whether *plausible* confounders have been ruled out will depend on careful examination of each case. For ailments that are responsive to expectations (such as depression and pain) the confounding effects of expectations will have to be ruled out, which can be achieved by blinding the patients and caregivers. When the assessment of outcomes is prone to influence from observer bias (such as blood pressure), potential confounding by variable measurements has to be ruled out, perhaps by standardizing the measurement procedure and by blinding the investigators in charge of collecting the data and evaluating the outcomes.

Yet sometimes the *strength* of the association (the size of the effect) will be greater than the combined effect of plausible confounders. In these cases, although plausible confounders have not been ruled by the design of the study, the large observed effect has swamped the combined effects of any plausible confounders. For example, the observed effects of general anaesthesia are unlikely to be accountable by selection bias, placebo effects or reporting bias. Thus, the failure to test the effects of general anaesthetics in double-blind, placebo controlled trials should not count against our beliefs that they cause reversible loss of consciousness.

Since one should compare the strength of association (size of effect) with the potential degree of bias, we have combined these into a single comparative guideline to emphasize this intrinsic comparison: *is plausible confounding less than the size of effect?*

A note of caution about strong *relative* effects (but small absolute effects) must be issued. Although 'weak' causes may be as real as 'strong' causes, it takes fewer (or 'weaker') confounders to account for a small absolute effect than for a large absolute effect. We therefore must be more careful when inferring from a strong relative (but small absolute) effect that an association is causal. At the same time, in many cases strong relative effects can provide strong support for the causal hypothesis. For instance, although the increased risk for lung cancer in smokers Bradford Hill cited was extremely low (0.07 per 1000 for non-smokers, 0.57 for smokers), the death rate for lung cancer in cigarette smokers was over 9 times the rate for non-smokers and thus provided good evidence for causation.[22]

Our omission of the 'experiment' guideline should not be interpreted as a sign that any observational study will do. Observational studies must demonstrate larger effects than randomized trials since they are at risk from selection bias (because the allocation to treatment groups is neither randomized nor concealed) and performance bias (because the participants and caregivers are not blinded). Whether the effect size in a particular observational study is sufficiently large to rule out the combined effects of selection and performance bias will vary from case to case. If investigators conducting an observational study have been vigilant in attempts to reduce selection bias (through careful selection of the control groups and *post hoc* adjustments), and the outcome is objective, the observational study might not have to demonstrate a dramatic effect in order to support causation.[25–27] In most other cases, however, the effect in an observational study will have to be dramatic in order to be confident that plausible confounders have been ruled out.[5]

In fact, our guideline can be more stringent than current EBM standards of evidence. According to hierarchies of evidence, RCTs with a low risk of bias often provide sufficient evidence to support causation. We require that, in addition to being at low risk, the effect size outweighs the combined effects of any residual bias. For example, although most systematic reviews of high quality RCTs of SSRIs suggest that these drugs enjoy a statistically significant benefit over 'placebo',[28,29] the absolute benefit is modest – a recent study suggests it is 6% (2–9%).[30] Yet one often overlooked source of confounding in these studies is the identifiable side-effects of the drug. If patients identify the drugs because of the side-effects (and independently of their effects on depression), then their expectations regarding recovery might be higher than if they knew they were taking a 'mere' placebo. To rule out the possible confounding effect of expectations, 'active placebos', which imitate the side-effects of SSRIs need to be employed. A systematic review of antidepressants versus 'active' placebos found that the drug less placebo difference was substantially reduced.[31] Besides confounding expectations, systematic reviews of SSRIs (like most systematic reviews) are likely to be confounded to some degree by publication bias,[32,33] funding source bias[34] and data mining in the original studies.[35] A careful calculation of the combined effects of these plausible confounders must be made before claiming that the systematic reviews of SSRIs support the claim that the drugs cause the reduction in depressive symptoms. Such calculations have not (to our knowledge) been made, so this guideline, unlike current hierarchies, does not necessarily support the existence of (non-placebo) effects of SSRIs.

Appropriate temporal and spatial proximity (encompassing and extending Bradford Hill's 'Temporality')

'Does a particular diet lead to disease or do the early stages of the disease lead to particular dietetic habits?'[22] The temporal part of this guideline is necessary: causes precede their effects and is therefore a true criterion. However, we should also ask: is the time *interval* between cause and effect consistent with the supposed mechanism? In general, the shorter the temporal and spatial interval, the less room for confounders (especially spontaneous remission) to interfere. It is equally important, for the time interval between administration of the treatment and cure to agree with the supposed mechanism of the treatment.

In some cases the *spatial* proximity between the site of administration and the outcome (see the oral

ulceration example below) may support causality – for example, thrombophlebitis at the site of injection of a cytotoxic drug. Again, the outcome need not be close to where the intervention was administered in order for the relationship to be causal, but spatial proximity generally leaves less room for confounders to interfere.

Dose responsiveness (Bradford Hill's 'Biological gradient')

Does the outcome change when the intensity of the intervention is altered (at least if the purported mechanism predicts such a relationship)? While the presence of a dose-response relationship does not always support causality (this guideline will not be applicable for 'all or none' causes), its absence *when expected* would lead us to doubt causality. Strongest 'dose-response' evidence comes when the process is reversible. For example, the risk of lung cancer is increased in smokers but is also reduced by a half in those who stop smoking at the age of 50 years and almost completely abolished in those who stop at the age of 30.[36]

Mechanistic evidence

Direct evidence does not always tell us *how* the intervention caused the outcome and this makes the result difficult to generalize.[37] What happens in between the intervention and the outcome is, as far as this category is concerned, a 'black box' (Figure 1). For example, Doll and Hill's famous study of the relation between the number of cigarettes smoked and the incidence of lung cancer[38] did not refer in any way to what happens between inhalation of cigarette smoke and the development of tumours in the lung. This brings us to the second category of guidelines.

Mechanisms play several roles. First, we tend to feel more confident about a treatment if the mechanism can be explained. Moreover, understanding the mechanism guides our generalization of a tightly controlled study to a wider population. Also, evidence about mechanisms plays a major role in generating hypotheses that should be tested by 'direct' tests. However, these roles of mechanism must be clearly distinguished from its distinct potential role in *confirming* hypotheses.

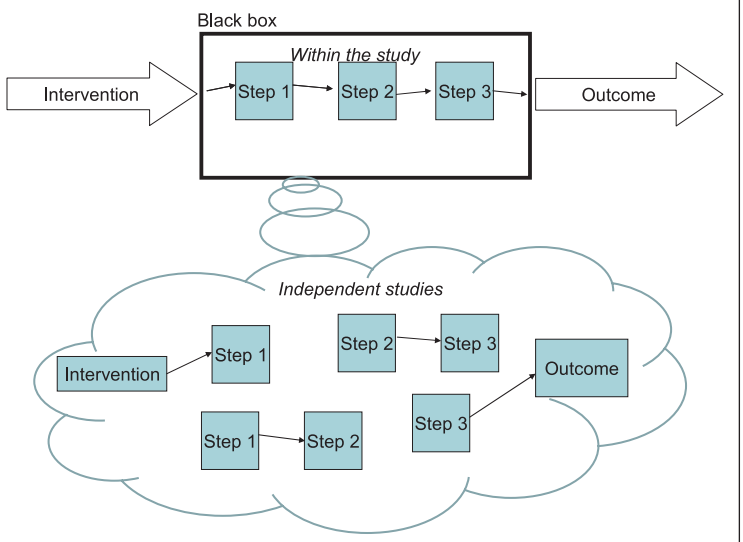
Although we believe that mechanistic evidence can provide evidential support for a causal hypothesis, two warnings are in order. Firstly, there is a difference between merely positing a mechanism (one can find a theory to explain almost anything) and providing sound evidence that there is a causal chain linking the intervention and the outcome. Secondly, appeal to mechanistic evidence has often justified the widespread use of treatments that turned out to be harmful.[40–46] Likewise, the *absence* of a plausible mechanism has often been used as a justification to ignore useful therapies such as antisepsis[47] and peptic ulceration.[48] With this in mind, although we believe that mechanistic evidence cannot be ignored, we acknowledge that mechanistic evidence should always play a subsidiary confirmatory role *vis-à-vis* direct evidence.

Plausible mechanism

Is there evidence supporting the causal chain linking the intervention and the outcome? For example, trials testing the effect of ACE inhibitors on reduction in stroke mortality might include evidence that ACE inhibitors reduce blood pressure, that reduced blood pressure reduces the risk of stroke, and that the reduced incidence of stroke reduces mortality. Of course, each 'step' in the causal process is a new 'black box'. For example, the link between ACE inhibitors and blood pressure can be further decomposed into a series of steps, until (in a reductionist model) we bottom out at the molecular level. Bradford Hill, no doubt as an oversight, implied that plausibility was limited to 'biological plausibility'. Mechanisms of action can also be mechanical (as in the Mother's Kiss example below) or chemical (as in the oral ulceration example below).

We can envisage three 'levels' of evidential support from mechanistic evidence. Firstly, the direct study can also include studies of the causal links between the intervention and the outcome (Figure 1, top half). A second level of mechanistic evidence is when the purported mechanism of action has been demonstrated in other, independent studies (Figure 1, bottom half). For example, separate studies could establish a probable link between ACE inhibition and lower blood pressure. Obviously, having evidence for a part of the mechanism is not as strong as evidence for all the links in the causal chain.

Figure 1
Direct evidence of probabilistic dependence of outcome on intervention + evidence for the causal process*



*Although Figure 1 illustrates the simple case in which the stages of the mechanism are linear, the relationship could be much more complex and include forks, cycles[39] and interactions

The second level of mechanistic evidence is closest to Bradford Hill's 'Coherence', and we have kept this guideline separate.

Coherence

Does the causal hypothesis *cohere* with what is currently known, or is it contradicted by current knowledge? This is best explained by what happens when the evidence does not cohere. For example, the causal process by which a homeopathic remedy is purportedly effective (other than by 'placebo' effects) is not currently explicable by mainstream science. Given the numerous examples where treatments that seemed to cohere with current science that turned out to be harmful,[40–46] and where treatments that seemed *not* to cohere with current science that turned out to be helpful,[47,48] this guideline must be applied with care.

Parallel evidence

There are rarely cases where there is only a single piece of evidence for a causal claim. When assessing whether an association is causal it is obviously necessary to consider *all* the relevant studies – this is the powerful idea underlying the importance of systematic reviews.

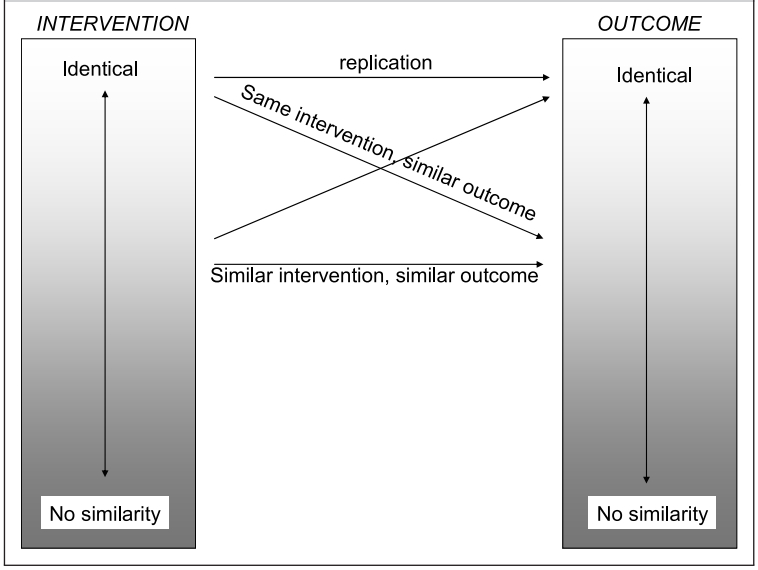
Replicability (Bradford Hill's 'Consistency')

A study can be replicated, which means that the same intervention is tested on a similar population, using the same outcome measure. In order to count as a replication, all the elements of the study must be kept constant as far as possible. Replicability is a central tenet of scientific method: if the experiment can be repeated and provides the same results, the chances that the original results arose due to confounding is reduced. If an experiment is not replicable, either something is wrong with the attempt to replicate it or the initial experiment must be questioned.

Similarity (of the study to other studies)

No two studies are absolutely identical, so similarities form a spectrum (Figure 2). Broadly speaking, there are several axes along which studies can differ. Firstly, the intervention can be different. If one NSAID reduced pain, we might have legitimately increased confidence that a new, similar drug would also reduce pain (although due caution would be warranted about potential adverse effects of the new drug and the benefit to harm balance). Other studies might use the same intervention and change the circumstances in which the intervention is administered. For example, we could test the intervention in a different (older or younger) population, conduct animal or *in vitro* experiments. We could also change the (geographical or socioeconomic) setting, or even the study type. Then, studies could use the same intervention but measure the outcome in different ways. If all the parallel studies gave similar results, then the causal hypothesis will be more strongly supported; if they don't, then we will have grounds to suspect either some of the parallel studies or the causal hypothesis itself. Of course, each piece of parallel evidence must be independently evaluated for validity (whether it satisfies the requirements inherent in our revised guidelines).

Figure 2
Types of similarities (the axis of 'similarity of circumstances' is omitted for simplicity)



Omitted guidelines

Besides *experiment*, which was absorbed in our first revised guideline, we also omitted *specificity*. Diseases usually have multiple causes and multiple effects, while most interventions also have multiple effects. In fact, Bradford Hill did not support this guideline with adequate examples, and in his description of multiple regression he admits that most diseases have multiple causes and that most causes have multiple effects.[22] For example, the fact that smoking increases the risk of lung cancer in no way repudiates evidence that smoking causes other diseases. Similarly, the fact that Prozac might have a positive effect on depression does not reduce the force of the claim that it also cures premature ejaculation.

Tests of whether the Revised Bradford Hill guidelines deliver the verdict of strong evidence for causation, even if RCTs have not been conducted

A strict application of the EBM evidence hierarchy would deliver the verdict that the following treatments are supported by relatively *poor* evidence since they have not been tested in randomized trials. After describing the examples, we shall evaluate whether the Revised Bradford Hill guidelines deliver a more reasonable verdict.

The Mother's Kiss

Glasziou *et al.*[5] cite the following example:

> *A child presented with a plastic bead lodged high in one nostril. The doctor asked for forceps, but the nurse suggested trying the mother's kiss technique – occluding the unblocked nostril while the mother blows into the child's mouth. The bead was thus easily dislodged and retrieved.*[5]

Most would agree that a single case (or at most a series of a few cases) would suffice to support claims that the mother's kiss caused the bead to dislodge.

Oral ulceration due to topical aspirin

Aronson and Hauben[49] have described several categories of adverse events related to drug administration that seem to require little more than anecdotal evidence to provide sufficiently strong evidence that the events are caused by adverse drug reactions. One of the categories is 'specific anatomical location or pattern of injury', in which:

> *… the location or pattern of injury is sufficiently specific to attribute the effect to the drug without the need for implicit judgment or formal investigation. The mechanism of injury can be related to physicochemical or pharmacological properties of the drug. Examples include extravasation reactions to cytostatic drugs and oral ulceration due to topical aspirin.*[49]

Here, anecdotal observations provide strong evidence that a particular drug caused an adverse event.

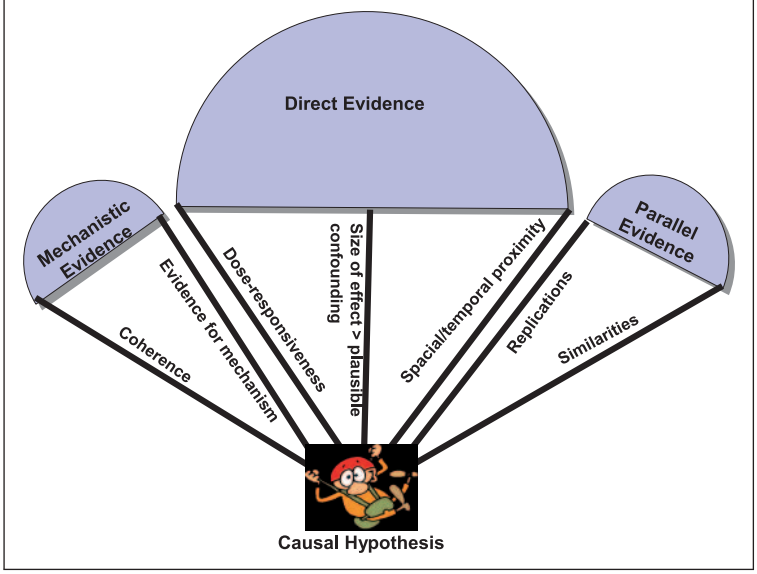
The Revised Bradford Hill guidelines deliver clear verdicts about the effectiveness of the Mother's Kiss and oral ulceration due to topical aspirin (Table 2). Admittedly the examples we chose are uncontroversial, but that is precisely why we chose them. Since nobody denies that these interventions caused their effects, while current hierarchies would deliver a poor 'grade' to their evidence base, it suggests that the Revised

Table 2
Applying the Revised Bradford Hill guidelines

		Mother's kiss	*Oral ulceration*
Direct	1. Size of effect not attributable to plausible confounding	Yes (dramatic effect; confounders highly unlikely)	Yes (dramatic effect; confounders highly unlikely)
	2. Appropriate temporal and/or spatial proximity	Yes (cure immediately follows the intervention and is spatially associated)	Yes (the effect is in immediate proximity to the intervention)
	3. Dose-responsiveness and reversibility	Not tested and not relevant (might have been tested by varying levels of expiratory force)	Not tested (dose-responsiveness not tested; but subsequent healing suggested reversibility)
Mechanistic	4. Plausible mechanism of action	Yes	Yes (acidic compound)
	5. Coherence	Yes (nothing contradicts the causal hypothesis)	Yes (nothing contradicts the causal hypothesis)
Parallel	6. Replicability	Yes	Not tested
	7. Similarity	Not relevant	Yes (aspirin causes gastric erosions)
Total		5 'yes' *(1, 2, 4, 5, 6)* 2 'not relevant' or 'not tested' *(3,7)*	5 'yes' *(1, 2, 4, 5, 7)* 2 'not relevant' or 'not tested' *(3, 6)*
VERDICT		*5 out of 7 guidelines satisfied*	*5 out of 7 guidelines satisfied*

guidelines can be useful tools for the future development and evolution of standards of medical evidence.

Figure 3
How different types of evidence support the causal hypothesis



Conclusions: suggesting ways to revise current hierarchies of evidence

The original Bradford Hill Guidelines can be simplified (some of the guidelines can be omitted while others can be combined or modified) and organized into three categories: *direct*, *mechanistic* and *parallel* evidence. In their revised form they suggest two ways that can inform revisions to current hierarchies of evidence. Firstly, it is more important for 'direct' evidence to demonstrate that the effect size is greater than the combined influence of plausible confounders, than it is for the study to be experimental. This view is compatible with the spirit of EBM hierarchies: the motivation for placing RCTs at the pinnacle of evidence hierarchies is that they generally rule out more confounders than other study types. If an observational study reveals an effect large enough to swamp the effects of any additional confounding then other study designs must be regarded as on a par with RCTs. Likewise, RCTs must demonstrate effect sizes sufficiently large to rule out the combined effect of any inevitable bias. Secondly,

the revised guidelines illustrate how different types of evidence can complement one another (Figure 3).[50,51] Whereas a trial is often open to the objection that it is an anomaly or not generalizable, if we supplement the evidence from the trial with strong mechanistic and parallel evidence, it becomes increasingly difficult to question the results of the study and its applicability to a wider target population. A similar idea supports the use of systematic reviews, teleoanalysis[33] and the tenet of replicability in scientific method. These features of the guidelines make them particularly helpful where RCTs are unfeasible.

References

1 Altman DG, Schulz KF, Moher D, *et al*. The revised CONSORT statement for reporting randomized trials: explanation and elaboration. *Ann Intern Med* 2001;**134**:663–94
2 Keen HI, Pile K, Hill CL. The prevalence of underpowered randomized clinical trials in rheumatology. *J Rheumatol* 2005;**32**:2083–8
3 Fergusson D, Glass KC, Waring D, Shapiro S. Turning a blind eye: the success of blinding reported in a random sample of randomised, placebo controlled trials. *BMJ* 2004;**328**:432
4 Hróbjartsson A, Forfang E, Haahr MT, Als-Nielsen B, Brorson S. Blinded trials taken to the test: an analysis of randomized clinical trials that report tests for the success of blinding. *Int J Epidemiol* 2007;**36**:654–63
5 Glasziou P, Chalmers I, Rawlins M, McCulloch P. When are randomised trials unnecessary? Picking signal from noise. *BMJ* 2007;**334**:349–51
6 Smith GC, Pell JP. Parachute use to prevent death and major trauma related to gravitational challenge: systematic review of randomised controlled trials. *BMJ* 2003;**327**:1459–61
7 Harbour RT, ed. *SIGN 50: A guideline developer's handbook.* Edinburgh: NHS Quality Improvement Scotland; 2008
8 Phillips B, Ball C, Sackett D, *et al*. Oxford Centre for Evidence-based Medicine Levels of Evidence. Oxford: CEBM; 2001. See http://www.cebm.net/index.aspx?0=1025 (accessed 30 January 2009)
9 Canadian Task Force on the Periodic Health Examination. The periodic health examination. *CMAJ* 1979;**121**:1193–254
10 US Preventive Services Task Force. *Guide to Clinical Preventive Services.* Washington, DC: US Department of Health and Human Services; 1996
11 Atkins D, Best D, Briss PA, *et al*. Grading quality of evidence and strength of recommendations. *BMJ* 2004;**328**:1490
12 Atkins D, Briss PA, Eccles M, *et al*. Systems for grading the quality of evidence and the strength of recommendations II: pilot study of a new system. *BMC Health Serv Res* 2005;**5**:25
13 Guyatt GH, Oxman AD, Vist GE, *et al*. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. *BMJ* 2008;**336**:924–6
14 Shang A, Huwiler-Muntener K, Nartey L, *et al*. Are the clinical effects of homoeopathy placebo effects? Comparative study of placebo-controlled trials of homoeopathy and allopathy. *Lancet* 2005;**366**:726–32
15 Williamson J, Russo F. Interpreting Causality in the Health Sciences. *Int Stud Philos Sci* 2007;**21**:157–70
16 Marshall T. Bradford-Hill Criteria provide the way ahead for controversial theory. *Int J Surg* 2005;**3**:287–8
17 Perrio M, Voss S, Shakir SA. Application of the bradford hill criteria to assess the causality of cisapride-induced arrhythmia: a model for assessing causal association in pharmacovigilance. *Drug Saf* 2007;**30**:333–46
18 Shakir SA, Layton D. Causal association in pharmacovigilance and pharmacoepidemiology: thoughts on the application of the Austin Bradford-Hill criteria. *Drug Saf* 2002;**25**:467–71
19 Staudenmayer H, Binkley KE, Leznoff A, Phillips S. Idiopathic environmental intolerance: Part 2: A causation analysis applying Bradford Hill's criteria to the psychogenic theory. *Toxicol Rev* 2003;**22**:247–61
20 Staudenmayer H, Binkley KE, Leznoff A, Phillips S. Idiopathic environmental intolerance: Part 1: A causation analysis applying Bradford Hill's criteria to the toxicogenic theory. *Toxicol Rev* 2003;**22**:235–46
21 van Reekum R, Streiner DL, Conn DK. Applying Bradford Hill's criteria for causation to neuropsychiatry: challenges and opportunities. *J Neuropsychiatry* 2001;**13**:318–25
22 Hill ABS, Hill ID. *Bradford Hill's principles of medical statistics.* 12th edn. Edinburgh: Edward Arnold; 1991
23 Hill AB. The Environment and Disease: Association or Causation? *Proc Roy Soc Med* 1965;**58**:295–300
24 Phillips CV, Goodman KJ. The missed lessons of Sir Austin Bradford Hill. *Epidemiol Perspect Innov* 2004;**1**:3
25 Concato J. Observational versus experimental studies: what's the evidence for a hierarchy? *NeuroRx* 2004;**1**:341–7
26 Benson K, Hartz AJ. A comparison of observational studies and randomized, controlled trials. *N Engl J Med* 2000;**342**:1878–86
27 Wood L, Egger M, Gluud L, *et al*. Empirical evidence of bias in treatment effect estimates in controlled trials with different interventions and outcomes: meta-epidemiological study. *BMJ* 2008;**336**:601–5
28 Arroll B, Macgillivray S, Ogston S, *et al*. Efficacy and tolerability of tricyclic antidepressants and SSRIs compared with placebo for treatment of depression in primary care: a meta-analysis. *Ann Fam Med* 2005;**3**:449–56
29 Williams JW Jr, Mulrow CD, Chiquette E, Noel PH, Aguilar C, Cornell J. A systematic review of newer pharmacotherapies for depression in adults: evidence report summary. *Ann Intern Med* 2000;**132**:743–56
30 Nemeroff CB, Entsuah R, Benattia I, Demitrack M, Sloan DM, Thase ME. Comprehensive analysis of remission (COMPARE) with venlafaxine versus SSRIs. *Biol Psychiatry* 2008;**63**:424–34
31 Moncrieff J, Wessely S, Hardy R. Active placebos versus antidepressants for depression. *Cochrane Database Syst Rev* 2004;**1**:CD003012
32 Whittington CJ, Kendall T, Fonagy P, Cottrell D, Cotgrove A, Boddington E. Selective serotonin reuptake inhibitors in childhood depression: systematic review of published versus unpublished data. *Lancet* 2004;**363**:1341–5
33 Barbui C, Cipriani A. Publication bias in systematic reviews. *Arch Gen Psychiatry* 2007;**64**:868
34 Perlis RH, Perlis CS, Wu Y, Hwang C, Joseph M, Nierenberg AA. Industry sponsorship and financial conflict of interest in the reporting of clinical trials in psychiatry. *Am J Psychiatry* 2005;**162**:1957–60
35 Procopio M. The multiple outcomes bias in antidepressants research. *Med Hypotheses* 2005;**65**:395–9

36 Doll R, Peto R, Boreham J, Sutherland I. Mortality in relation to smoking: 50 years' observations on male British doctors. *BMJ* 2004;**328**:1519

37 Cartwright N. *Hunting causes and using them : approaches in philosophy and economics.* Cambridge: Cambridge University Press; 2007

38 Doll R, Hill AB. The mortality of doctors in relation to their smoking habits; a preliminary report. *BMJ* 1954;**1**:1451–5

39 Machamer P, Darden L, Craver CF. Thinking About Mechanisms. *Philos Sci* 2000;**67**:1–25

40 Piaggio G, Elbourne DR, Altman DG, Pocock SJ, Evans SJ. Reporting of noninferiority and equivalence randomized trials: an extension of the CONSORT statement. *JAMA* 2006;**295**:1152–60

41 Echt DS, Liebson PR, Mitchell LB, *et al*. Mortality and morbidity in patients receiving encainide, flecainide, or placebo. The Cardiac Arrhythmia Suppression Trial. *N Engl J Med* 1991;**324**:781–8

42 Takala J, Ruokonen E, Webster NR, *et al*. Increased mortality associated with growth hormone treatment in critically ill adults. *N Engl J Med* 1999;**341**:785–92

43 Hayes MA, Timmins AC, Yau EH, Palazzo M, Hinds CJ, Watson D. Elevation of systemic oxygen delivery in the treatment of critically ill patients. *N Engl J Med* 1994;**330**:1717–22

44 Hebert PC, Wells G, Blajchman MA, *et al*. A multicenter, randomized, controlled clinical trial of transfusion requirements in critical care. Transfusion Requirements in Critical Care Investigators, Canadian Critical Care Trials Group. *N Engl J Med* 1999;**340**:409–17

45 Rossouw JE, Anderson GL, Prentice RL, *et al*. Risks and benefits of estrogen plus progestin in healthy postmenopausal women: principal results From the Women's Health Initiative randomized controlled trial. *JAMA* 2002;**288**:321–33

46 Spock B, Fox D. *Baby and Child Care.* New York, NY: Pocket Books; 1966

47 Gillies D. Hempelian and Kuhnian approaches in the philosophy of medicine: the Semmelweis case. *Stud Hist Philos Sci C Stud Hist Biol Biomed Sci* 2005;**36**:159–81

48 Marshall B. Helicobacter connections. *Chem Med Chem* 2006;**1**:783–802

49 Aronson JK, Hauben M. Anecdotes that provide definitive evidence. *BMJ* 2006;**333**:1267–9

50 Wald NJ, Morris JK. Teleoanalysis: combining data from different types of study. *BMJ* 2003;**327**:616–18

51 Aronson JK. Unity from diversity: the evidential use of anecdotal reports of adverse drug reactions and interactions. *J Eval Clin Pract* 2005;**11**:195–208