This article was downloaded by: [Australian National University Library]

On: 4 May 2009

Access details: Access Details: [subscription number 907447645]

Publisher Routledge

Informa Ltd Registered in England and Wales Registered Number: 1072954 Registered office: Mortimer House, 37-41 Mortimer Street, London W1T 3JH, UK

Social Epistemology

Publication details, including instructions for authors and subscription information:

<http://www.informaworld.com/smpp/title-content=t713765921>

A Couple of the Nasties Lurking in Evidence-Based Medicine

Jason Grossman

Online Publication Date: 01 October 2008

To cite this Article Grossman, Jason(2008)'A Couple of the Nasties Lurking in Evidence-Based Medicine',Social Epistemology,22:4,333 — 352

To link to this Article: DOI: 10.1080/02691720802566961

URL: <http://dx.doi.org/10.1080/02691720802566961>

PLEASE SCROLL DOWN FOR ARTICLE

Full terms and conditions of use: <http://www.informaworld.com/terms-and-conditions-of-access.pdf>

This article may be used for research, teaching and private study purposes. Any substantial or systematic reproduction, re-distribution, re-selling, loan or sub-licensing, systematic supply or distribution in any form to anyone is expressly forbidden.

The publisher does not give any warranty express or implied or make any representation that the contents will be complete or accurate or up to date. The accuracy of any instructions, formulae and drug doses should be independently verified with primary sources. The publisher shall not be liable for any loss, actions, claims, proceedings, demand or costs or damages whatsoever or howsoever caused arising directly or indirectly in connection with or arising out of the use of this material.

A Couple of the Nasties Lurking in Evidence-Based Medicine

Jason Grossman

The Evidence-Based Medicine (EBM) movement is an ideological force in health research and health policy which asks for allegiance to two types of methodological doctrine. The first is the highly quotable motherhood statement: for example, that we should make conscientious, explicit and judicious use of current best evidence (paraphrasing Sackett). The second type of doctrine, vastly more specific and in practice more important, is the detailed methodology of design and analysis of experiments. This type of detailed methodological doctrine tends to be simplified by commentators but followed to the letter by practitioners.

A number of interestingly dumb claims have become entrenched in prominent versions of these more specific methodological doctrines. I look at just a couple of example claims, namely:

1. *Any randomised controlled trial (RCT) gives us better evidence than any other study.*
2. *Confidence intervals are always useful summaries of at least part of the evidence an experiment gives us about a hypothesis.*

To offer a positive doctrine which might move us past the current conflict of micro-theories of evidence, I propose a mild methodological pluralism: in any local context in which none of a variety of scientific methodologies is clearly and uncontentionously right, researchers should not be discouraged from using any methodology for which they can provide a good argument.

Keywords: *Evidence; Evidence-Based Medicine; Methodology; Statistics; Pluralism*

Jason Grossman is a Lecturer in the School of Humanities, Australian National University. He has worked in both health policy and biostatistics. Correspondence to: Jason Grossman, School of Humanities, A. D. Hope Building, Australian National University, Acton ACT 0200, Australia. Email: Jason.Grosman@anu.edu.au

What is Evidence-Based Medicine?

Evidence-Based Medicine (EBM) is an ideology (by which I mean a simple idea with an associated political programme—something which is no bad thing in itself) which originally stood for the extension of certain parts of the methodology of clinical epidemiology, with its strong emphasis on biostatistics, to the whole of the practice of clinical medicine. Then, from the late 1980s, EBM gradually became the view that those methods ought to be applied to all medical and health-related research and (importantly) health policy. For details of this history, see Daly (2005).

The most famous definition of EBM is the following, by Sackett, one of the two or three people most responsible for the current form of the EBM movement:

Evidence-based medicine is the conscientious, explicit and judicious use of current best evidence in making decisions about the care of individual patients. (Sackett et al. 1996)

This definition is inaccurate in at least two important respects. First of all, EBM is and almost always has been, since very shortly after its invention, applied to populations as well as individual patients and to health policy as well as patient care. In this respect, the definition may have been almost accurate when first suggested but is accurate no longer.

Secondly, while it is true that the EBM literature includes exhortations to use the best evidence, it is dominated not by this general rule but instead by a number of *extremely specific* rules. These rules are to be applied rigidly, as I will illustrate below, without any consideration in particular contexts of whether they fulfill Sackett's definition by helping us to use current best evidence. The EBM movement has justified to its own satisfaction (not to mine) the idea that its specific rules will help us to use current best evidence; but having done that, it tells us not to worry about the matter any further, and certainly not to evaluate its rules on a case-by-case basis. It gives us a Catholic doctrine of trusting the Church's rules, without the amelioration of the Catholic doctrine that we may also look to our consciences to judge the rules on a case-by-case basis. We will see this in action in the examples which I will present later.

Despite its inaccuracy, Sackett's definition is widely accepted, presumably on the grounds that one of the main originators of EBM ought to know what it is. It's worth pausing here to recall that Berkeley was tempted to define his subjective idealism as the merest common sense as held by the man in the street. Just as Sackett's definition embodies what Sackett believes to be the most salient facts about EBM, such a definition of subjective idealism as common sense embodies what Berkeley took to be the most salient facts about his theory. But it is clearly no good; at least, not as a definition. A much better definition of subjective idealism would be one which asked more specifically what we had to believe in order to believe it: such as, that objects are bundles of sensations. *Then* we could start arguing over whether it's true.

Similarly, Sackett's definition of EBM is neither good lexicography nor good history but merely Sackett's view of EBM's merits. Just as in the case of subjective idealism, a better definition would list the things that EBM tells us to believe, not in terms of vague success words like "best" but in terms of EBM's specific injunctions.

Unfortunately, EBM is a rag-bag of methodological imperatives with no unifying principle. A proper definition, taking into account everything which EBM actually stands for, would be gut-witheringly, boringly long, and excruciatingly disjunctive. So I will not give a definition. Instead, I will concentrate on two parts of EBM, based on the claims listed above. These are claims which would form part of any realistic definition.

Dumb Claim Number 1

Dumb claim of EBM number 1: *Any randomised controlled trial (RCT) gives us better evidence than any other study.*

Dumb claim 1 is not mentioned anywhere in the motherhood statements of EBM; but it is a central claim of *operationalised* versions of EBM, even when they are presented in the same publications as the motherhood statements.

EBM is operational through and through, because one of its central tenets is that “Systematic and explicit methods of making judgments can reduce errors and improve communication” (GRADE working group 2004, 1490). Consequently, the methods it recommends are more or less rigid and algorithmic. One of the most accessible ways in which the concepts of EBM have been operationalised is through evidence hierarchies. The EBM movement has produced a number of these, most of which are minor variants on each other. I will present four. For simplicity and brevity, I will start by concentrating on a particularly influential hierarchy produced by the US Preventive Services Taskforce. As well as being influential, the US Preventive Services Taskforce hierarchy is also typical.¹ Figure 1 shows the US Preventive Services Taskforce hierarchy alongside two Australian hierarchies, one of which I have analysed in Grossman and Mackenzie (2005). At the end of this section, I will briefly consider a more recent and more complex evidence hierarchy. I will show that it still upholds EBM’s dumb claim 1.

The EBM movement actively promotes these evidence hierarchies (more or less interchangeably), and the terminology of “level 1 evidence” has long ago become entrenched in the language of medical research, as illustrated in this extract from the minutes of the National Health and Medical Research Council, the Australian government’s most senior health policy committee:

The draft document [on acute pain management] was introduced to the Council by Professor Michael Cousins, who chaired the working group which prepared the document. ... Professor Cousins reported that much of the evidence within the document relates to the efficacy of one form of treatment versus another, there often being options. He also noted the general lack of level I evidence. (National Health and Medical Research Council 1998)

By implication, had level 1 evidence been available it would have been taken to exclude or override other evidence. The analysis in Grossman and Mackenzie (2005) gives more detailed reasons to think that this is the case. The ways in which the EBM literature recommends the hierarchies be used varies, and of course the ways in which they *are* used is another thing entirely. To document either or both of those exhaustively would be a very major undertaking which I have not undertaken fully; but to the extent that I have (both formally in Grossman and Mackenzie (2005) and informally), there

US Preventive Services Task Force (Preventive Services Task Force 1996)	MERGE (New South Wales Department of Health) (Liddle et al. 1996)	Medicare Services Advisory Committee (Medicare Services Advisory Committee 2000)
<p>Level I: evidence obtained from at least one properly randomised, controlled trial.</p> <p>Level II-1: evidence obtained from well-designed controlled trials without randomisation.</p> <p>Level II-2: evidence obtained from well-designed cohort or case-control analytic studies, preferably from more than one center or research group.</p> <p>Level II-3: evidence obtained from multiple time series with or without the intervention. Dramatic results in uncontrolled experiments (such as the results of the introduction of penicillin treatment in the 1940s) could also be regarded as this type of evidence.</p> <p>Level III: opinions of respected authorities, based on clinical experience; descriptive studies and case reports; or reports of expert committees.</p>	<p>Level I: systematic review of all relevant randomised controlled trials; large multi-center randomised controlled trials</p> <p>Level II: one or more randomised controlled trials and studies</p> <p>Level III: controlled trials without randomisation; cohorts; case-control analytic studies; multiple time series; before and after studies (preferably from more than one center or research group)</p> <p>Level IV: other observational studies</p>	<p>Level I: evidence obtained from a systematic review of all relevant randomised controlled trials</p> <p>Level II: evidence obtained from at least one properly designed randomised controlled trial</p> <p>Level III-1: evidence obtained from well-designed pseudo-randomised controlled trials (alternate allocation or some other method)</p> <p>Level III-2: evidence obtained from comparative studies with concurrent controls and allocation not randomised (cohort studies), case-control studies, or interrupted time series with a control group</p> <p>Level III-3: evidence obtained from comparative studies with historic control, two or more single-arm studies, or interrupted time series without a parallel control group</p> <p>Level IV: evidence obtained from case-series, either post-test or pre-test and post-test</p>

Figure 1 Sample Evidence Hierarchies.

seems to be consensus that the whole point of the hierarchies is to represent strict orderings of study types. Recall the point I made earlier that the EBM enterprise is largely algorithmic (operationalised), and deliberately so, so that decisions about (for example) which papers are worth reading can be automated. The strict separation of study types is part of this decision process.

Dumb Claim 1 is False

This brings us back to dumb claim 1, because dumb claim 1 is equivalent to the strict separation between the top levels of the evidence hierarchy.

What I want to show in the rest of this section is that dumb claim 1 is false. It is *not* the case that any randomised controlled trial (RCT) gives us better evidence than any other study.

The first thing we might note about dumb claim 1 is how strange it is to claim anything about how good the evidence from a study is based solely on the study's *design*.

A study can be designed well but just happen to give useless results, or vice versa; and yet dumb claim 1 contradicts these facts. So dumb claim 1 is in trouble right off the bat. It seems to me that that is a knock-down argument against dumb claim 1. I could stop there. But instead I will put that point to one side, because dumb claim 1 is going to get itself into trouble in quite a number of other ways.

The rest of my strategy will be something like this. I will show that a particular bad RCT was bad, and hence that bad RCTs can be bad. That is perhaps uncontentious. Then I will compare my bad RCT with a good study which was not an RCT. That will show pretty conclusively, I hope, that it is possible for evidence from level 1 of the evidence hierarchy to be worse than evidence from lower down in the hierarchy, *contra* dumb claim 1.

Each step in this argument will be blindingly obvious, and admitted to be so in the EBM literature whenever particular cases are considered; and yet the conclusion of the argument is contradicted by dumb claim 1, the evidence hierarchy continues to be taken seriously (and rigidly), and policymakers continue to prefer level 1 evidence without enquiring into the quality or outcomes of particular studies. Go figure.

All that remains is to look at a bad RCT and a good non-RCT.²

A Bad RCT

Patients with HIV infection were randomized to received [sic] either a "real" patient held record or a "control" patient held record and prospectively followed. Twenty of the 41 patients randomized to the study had records available for analysis. Eight of the 20 available patients were randomized to the real records and 12 to control records. Both groups were well matched. (Meese, Vujovic, and Fairley 1997, 226)

In this study, 41 HIV-positive people attending a single hospital clinic were given a record which either did (intervention) or didn't (control) hold a summary of their medical record, which they used to facilitate their interactions with the many medical

professionals they subsequently dealt with. The literature suggested that the intervention group should have had improved health outcomes.

This was a bad RCT. One reason it was bad was that it was small. (I will come to why small RCTs are bad in a moment.) Contrary to what one might wish, small RCTs are common. This is for three main reasons: lack of funding, wishful thinking about the outcome, and confusion about the level of analysis. To start with lack of funding: first of all, many RCTs are student projects, which are essentially self-funded. Secondly, while *medical* funding structures encourage researchers to fund projects “properly” or not at all, the same is not true in all disciplines. That’s all I can say about funding; much more could of course be said, but not briefly.

To turn to wishful thinking about the outcome: practically all studies have their sample sizes chosen by “power calculations” which depend on the expected outcome. But the expected outcome is often chosen by the experimenters themselves, who have a vested interest in running an affordable study. Consequently, the sample size can be, and presumably often is (and in my own experience certainly is), chosen to be precisely that value which allows the study to be run, even if this makes the study too small to really be useful. (Admittedly, constraints such as peer review can complicate this story.)

Finally, and most interestingly from the social point of view, there is the possibility of confusion about the level of analysis (Evidence-Based Medicine Working Group 2002). I’ll illustrate this first with a hypothetical example. Imagine an anti-drug campaign in high schools. The evaluation of the campaign is an RCT which enrolls 6,000 school students, 600 of whom are in each of 10 schools. What is the sample size? 6,000? For some purposes, yes. But for the purposes that matter most, the sample size is 10, because we only sampled 10 schools. The point here is that the sampling which matters most is the sampling across the main type of variation that we’re interested in. After all, we would not say that the sample size was 24,000 on the grounds that we had sampled 24,000 students’ limbs; and that’s because the variation between students’ limbs is of no interest in this context. When there’s a social variable of interest (as of course there often is), it’s *that* variable which should determine the sample size of most interest. In this school campaign the variable of by far the most interest is the school.³ And in terms of that variable we’ve only sampled 10 things.

This issue crops up in the real RCT by Meese, Vujovic, and Fairley. The claimed sample size was 20; but that was the number of patients. What Meese et al. were studying was the effectiveness of having the patients carry around summarised versions of their medical records. So there was a very important social variable, namely the local medical systems those patients found themselves in. The real sample size, in terms of that variable, was plausibly the number of hospitals studied. In terms of that variable, the sample size was one.

Even if the sample size was 20, as claimed, that was still rather small, so we had better look at the well-known reason why small randomised studies are bad. The reason is that *randomising, in a small study, is likely to be either pointless or counterproductive*. This is most obvious with a really tiny study (although it’s also true with a merely small study like Meese’s). Take a study with just four people in it, with an intervention and a control group, and suppose that two of the people are male and two female. Are you

going to randomise the people into the groups? If you do, there's a 50% chance that both males will be in one group and both females in the other. If two of them speak only English and two are bilingual, there's a 50% chance that you'll randomise both bilingual speakers into one group and both monolingual speakers into the other. And so on. You'd have a much better chance of balancing these variables if you did it deliberately, by choosing your groups, than if you left it to randomisation. Maybe you should randomise anyway, in a desperate attempt to balance variables you don't know about. But really you'd be much better off abandoning the study.

This illustrates that in a small RCT the intervention and control groups are likely to be wildly mismatched. When that is the case, it's unreasonable to draw any conclusions at all from the study (although one can amuse oneself by drawing conclusions about the authors instead or, more fairly, about their research culture).⁴

I'm claiming that the intervention and control groups in small RCTs are likely to be wildly mismatched. How can I square this with the statement by Meese et al. that "Both groups were well matched"? Perhaps surprisingly, this is easily done. Epidemiologists like Meese et al. are allowed to make claims of that sort when they've made significance tests of the main variables against the hypothesis that the intervention and control groups are exactly the same. There are two problems with that. (As usual in this paper, these are problems well known to theoretical statisticians.) First of all, we are not interested in whether the intervention and control groups are exactly the same, but rather in whether they are wildly different. Secondly, the very same small sample size which caused me to go into conniptions in previous paragraphs causes these significance tests to be extremely weak, and therefore to fail to reject almost any hypothesis. Putting these two technicalities together, here is what we get. Meese et al.'s significance tests failed to reject the hypothesis that the two groups were exactly the same; they would also (I bet) have failed to reject the hypothesis that they were wildly different.⁵ This illustrates the fact that the standard statistical tests to see whether small groups of subjects are well matched are completely meaningless.⁶

That summarises the reasons why small RCTs are bad. And there are other reasons why the study by Meese et al. is a bad RCT. One is that it is a social intervention treated as if it were an individual intervention. Almost whenever this happens, we should expect the effects of the study to "bleed" from the intervention group into the control group. In a little more detail: this is a social intervention because it has its effects (if at all) by affecting a whole social system, consisting not just of a patient but also of a number of medical professionals in interaction with the patient and with each other. If that were not the case, then the patient-held record would not be efficacious. (At least, this is a plausible theory of the basis on which one might expect the intervention to work.) Whatever the effect on the social system is, it may be presumed to work in favour of the people in the control group too, thus (a) distorting the results and (b) weakening any beneficial effects of randomising and further weakening dumb claim 1.

We have not yet finished with Meese et al. Note that they started with 41 subjects but ended up with only eight in the intervention group and 12 in the control group, so there was very substantial dropout from the study. The most interesting and worrying aspect of this dropout is that the analysis of the results treated this dropout as if it

occurred evenly between the intervention and the control groups. But if the dropout from an RCT is substantially different between the groups (note: *not* if it's statistically significantly different, which in a small trial it almost never will be, but if it's substantially different) and if the difference is inexplicable (which it almost always is) then all bets are off: the study is a goner. The problem is somewhat hidden (at least to my eye) by the small sample size, which means that one group has only four more people than the other; but those four people represent a huge proportion of the sample, so the dropout really was substantially different between the two groups, for some unknown reason. But the results of the study consist *entirely* of a comparison between the two groups. And that comparison may claim to be telling us something about the intervention ... but really (for all we know or have any right to guess) it's telling us just as much about whatever it was that caused the differential dropout.

So much for an example of a bad RCT, which, despite being bad, was still an RCT, and therefore produced level 1 evidence.

Now let's look at a study which fits into the evidence hierarchy much lower down, giving us only level 2 evidence—according to the hierarchy, much less reliable stuff.

A Good Non-RCT

MacGowan et al. (1997) studied clients of seven methadone maintenance clinics in two States in New England (northeastern USA) in the 1990s. These were people who were being given the drug methadone in the hope that it would reduce their dependence on heroin; approximately 20% were HIV positive. MacGowan et al. wanted to know what the effects of the clinics' counseling and treatment programmes were on the clients' health behaviours, and they wanted to find this out prospectively, by conducting interviews which tracked individual clients over a period of a year. This was an ambitious study, since people who are or have been addicted to heroin are often unwilling to be tracked by researchers over time.

MacGowan et al.'s study was not an RCT. In terms of the US Preventive Services Task Force hierarchy, it *cannot* provide level 1 evidence, and this is something we could know about it *beforehand*, no matter how well the study turned out. (Indeed, it can't even provide evidence at level II-1.) And yet it provides excellent evidence, much better than the results of the bad RCT.

The first interesting difference between the two studies, given my criticisms of the bad RCT we saw earlier, is that MacGowan et al. studied 1,053 people.⁷ Indeed, MacGowan et al.'s study seems to have been a response to previous studies which were too small (Calsyn et al. 1982; Landis, Earp, and Koch 1992).

The second interesting difference is that MacGowan et al. obtained detailed ethnographic information from their study subjects in their interviews, including demographic, sexual, and health information relevant to HIV. They were able to match this with clinical records for many subjects. This enabled them to analyse their study results according to a variety of important factors which could have skewed the outcomes. Contrast this with Meese et al.'s simple claim that their intervention and control group were well-matched which, as we have seen, was meaningless.

Neither of these benefits of MacGowan et al.'s study over Meese et al.'s earns them any credit in terms of the evidence hierarchy.

MacGowan et al. discovered that people who remained in methadone treatment and counselling were decreasingly likely, over time, to:

- inject illegal drugs at all;
- inject with dirty equipment; and
- have sexual partners without condoms.

Importantly, these results were obtained using multiple regression, which made some (admittedly fallible) attempt to take into account all the relevant social variables which MacGowan et al. (but not Meese et al.) had measured.

So, to summarise this section and the previous one: we've seen a bad RCT which didn't teach us anything, and we've seen a good study which wasn't an RCT but which taught us a lot. Perhaps not surprisingly, dumb claim 1 turned out to be false.

Does EBM Really Support Dumb Claim 1?

Selective quoting from the vast EBM literature, like selective quoting from the Bible and Shakespeare, can be made to support almost any conclusion. Having shown how silly dumb claim 1 is, one might reasonably guess that I would have to have been selectively quoting in order to make it seem like anyone would ever have supported it.

But unlike Shakespeare (if not the Bible), the EBM literature reaches consensus about a number of things. This is so on a number of levels (most of which I cannot investigate here). One way in which the literature forms a consensus on dumb claim 1 is by returning to it repeatedly over time. I will illustrate this here with a slightly more recent evidence hierarchy.⁸

The Oxford Centre for Evidence-Based Medicine (OCEBM) is perhaps the most thoughtful and careful large group of EBM scholars. It has produced one of a number of evidence hierarchies which improve on the first generation shown above (See Figure 2).

The column headed "economic and decision analyses" is perhaps the one which should interest us the most. But it turns out that the studies which are praised by that column are those which combine economic analyses with "SR's" (systematic reviews)⁹ of level 1 studies. Looking across the page, we see that level 1 studies (for the OCEBM) are RCTs of a special type, or alternatively a very rare study called "All or none" which produces an effect in literally 100% of people exposed to an intervention and literally 0% of people denied the intervention.¹⁰ To all intents and purposes, then, level 1 studies are what they were in the simpler hierarchy we've been using all along, and the best economic and decision analyses are level 1 studies with economic analyses tacked on. Almost nothing has changed (although to be sure, what little has changed has changed for the better).

The main change (for our purposes) that's taken place between the first-generation evidence hierarchies and the OCEBM hierarchy is the increase in columns, from one column (all studies) to five columns, dealing separately with studies about "therapy/

Level	Therapy/Prevention, Aetiology/Harm	Prognosis	...	Economic and decision analyses
1a	SR (with homogeneity*) of RCTs	SR (with homogeneity*) of cohort studies; different populations		SR (with homogeneity*) of Level 1 economic studies
1b	Individual RCT (with narrow Confidence Interval†)	Individual inception cohort; 80% follow-up; population		Analysis based on clinically sensible costs or alternatives; systematic review(s) of the evidence; and including multi-way sensitivity analyses
1c	All or none§	All or none case-control		Absolute better-value or worse-value analyses ††††
2a	SR (with homogeneity*) of cohort studies	SR (with homogeneity*) of retrospective cohort studies; control groups		SR (with homogeneity*) of Level >2 economic studies
	.			.
	.			.
	.			.

Figure 2 The OCEBM Hierarchy (excerpt).
Source: Centre for Evidence-Based Medicine, undated.

prevention/aetiology/harm”, “prognosis”, “diagnosis”, “differential diagnosis/symptom prevalence study”, and “economic and decision analyses”. The diagnosis and differential diagnosis/symptom prevalence columns are similar (for our purposes) to the prognosis column.

Note the medical, clinical focus in the OCEBM terminology (as compared to the earlier terminology of the US Preventive Services Taskforce). The clinical terminology might make it seem at first glance as though the OCEBM hierarchy is limited to hospital interventions. This is not the case: almost every study that’s of interest in the broader health field is covered by the OCEBM table ... specifically, by the first and last columns, which are (as I have just shown) still committed to dumb claim 1 (except for the *very* minor caveat about all-or-nothing studies and the improvement in sometimes recognising sample size issues).¹¹

The clinical focus of the OCEBM table suggests an important point. It would be much easier to argue for an analogue of dumb claim 1 in a very limited clinical context, such as in large drug trials conducted entirely in non-ambulatory patients (where the drawbacks of bad RCTs discussed above are limited because social covariates are *likely* to be *relatively* unimportant, effects are *unlikely* to bleed from the intervention group into the control group, and so on). I would like to thank an anonymous referee of this paper for the idea that when EBM advocates make dumb claim number 1 they actually intend only to make some such narrower claim. Unfortunately, regardless of what they intend, that is not what they say; and, more importantly, they draw *inferences* from dumb claim 1 which belong to the broad claim which I am examining in this paper rather than to the hypothetical narrower one.¹²

One final change which has occurred in the short time since the first generation of evidence hierarchies can be seen in level 1b of the OCEBM evidence hierarchy. Here we see an explicit mention of RCTs which produce narrow confidence intervals. This leads nicely on to dumb claim 2.

Dumb Claim Number 2

Dumb claim of EBM number 2: *Confidence intervals are always useful summaries of at least part of the evidence an experiment gives us about a hypothesis.*

As we saw at the beginning of this paper, at one level of doctrine the EBM movement does not admit to believing in anything much (only “the conscientious, explicit and judicious use of current best evidence in making decisions about the care of individual patients”). But there is a level of nitty-gritty at which EBM really does stand for something. This paper is about a couple of the things it really stands for.

At the nitty-gritty, operational level, EBM stands for the application of certain specific statistical methods. In the eyes of its founding fathers, this presumably does not contradict the idea that we should use the best evidence, since presumably they think that only those specific statistical methods can give us the best evidence. However, there is no more reason to think that they are right about that than there would be to think that any other small group of doctors had once and for all established god’s own truth about the best set of specific statistical methods to use for all purposes.

Here I will examine one particularly important set of research methods: confidence intervals. These are recommended by EBM proponents as god’s own method of parameter estimation and hypothesis testing.

The EBM literature is full of injunctions to use confidence intervals under all circumstances. And yet it has been known to theoretical statisticians since long before the birth of EBM that confidence intervals sometimes give results which are known *in advance* to be completely useless.

Claims That We Should Use Confidence Intervals

The consolidation of the EBM movement in the late 1980s coincided with a major consolidation among the world’s most influential biostatisticians. Almost all at once (in an act of unanimity which is not surprising given the close-knit nature of the international biostatistical community), the major medical journals published papers and editorials asking medical researchers to replace their previous methods of measuring statistical uncertainty (all based on significance tests) by confidence intervals (Langman 1986; Gardner and Altman 1986; Berry 1986; Anonymous 1987). They reasoned that confidence intervals, which associate a pair of numbers with each result instead of the single number given by a significance test, allowed one to estimate the size of a result at the same time as deciding on its statistical significance.¹³

The recommendation was by no means new (confidence intervals are very nearly as old as the significance tests they replace (Neyman 1937)), but confidence intervals had not been very widely adopted up until that point; and so the concerted attempt by journals to make their authors use confidence intervals was something of a watershed. The EBM movement absorbed the new consensus, and made it a *sine qua non* of “best practice”. Since then, the battle for confidence intervals has been more or less won in the mainstream EBM literature, and continues in a set of ever-widening disciplinary

ripples: for example, it is currently being fought in the field of forecasting studies (Armstrong 2007).

The papers cited above introduced confidence intervals to the bulk of medical researchers, either directly or indirectly: they were paraphrased and cited by less prestigious journals as the innovation diffused. (Gardner and Altman (1986) has been cited at least 790 times.) Perhaps importantly, they failed to instill a good understanding of what confidence intervals are in their readers, starting with their very own journal editors. One of the first and presumably most influential editorials, in the *British Medical Journal*, went so far as to give an interpretation and implicit definition of confidence intervals which owes more to what we'd *like* a confidence interval to be than to what one is:

...consider a difference in blood pressure between groups of diabetics and non-diabetics of 6 mm Hg with a 95% confidence interval of 1.1 to 10.9 mm Hg.

...we also say using the confidence interval that *there is only a 2.5% chance that the true difference in the population at large is greater than 10.9 mm Hg*. ...the conclusion must be that we are unlikely to be missing a large and clinically important difference. (Langman 1986, 716; emphasis added)

Assigning probabilities to hypotheses like that is something which all theoretical statisticians agree confidence intervals can't do. We will see below an illustration of why they can't (at least, an illustration of one of several reasons why they can't). You may wonder what else statistical methods could possibly be for if not to allow us to assign probabilities to hypotheses. Wonder away; parts of the statistical literature wonder with you (notably Savage and discussants 1962). In the meantime, such assignments cannot be made using confidence intervals, and were never meant to be (Neyman 1937; Stuart, Ord, and Arnold 1999).¹⁴

Confidence Intervals Don't Always Work

In this section, I will use a toy example to show that confidence intervals don't always work. Toy examples are pernicious, of course. But this is a rare case in which the point that needs to be made is purely mathematical, so a toy example is relatively harmless. This example has been well known to theoretical statisticians for many decades. It shows how there can be a perfectly formed confidence interval which makes no sense.¹⁵

Here is the toy example. Suppose we need to write a protocol for the following common scenario. A patient is admitted to hospital having swallowed three toy academics. A machine that goes beep tells a surgeon that the three toys are stuck in a 2 cm length of the patient's colon. The surgeon would like to operate to remove the three toys. But there is a constraint. One of the toys is a toy Dean. This one has to be removed first. It's guarded on both sides by toy social epistemologists which (fortuitously) protect the colon from toxic paint which flakes off the toy Dean. If the surgeon accidentally removes one of the toy social epistemologists first the patient will die and the surgeon will be struck off. There's a 50% chance that the surgeon can feel the position

of one of the social epistemologists by palpation of the colon, and a 50% chance that she can feel the positions of both of them. She definitely can't feel the Dean.

Assuming the three toys are equally spaced with the Dean in the middle, we can get a 75% confidence interval for the position of the Dean,¹⁶ where x_1 and x_2 are the positions at which the surgeon felt something:

$$\begin{aligned} & x_1 - 1 \text{ to } x_1 + 1 \text{ if } x_1 = x_2 \\ & (x_1 + x_2) / 2 \text{ to } (x_1 + x_2) / 2 \text{ otherwise} \end{aligned}$$

Why is this a 75% confidence interval? First of all consider the first line of the confidence interval, and suppose that the surgeon could only feel one toy, at position x_1 . There's a 50% chance of this happening. Since we're assuming that the toys are equally spaced over 2 cm, with the Dean exactly in the middle, the Dean has to be at $x_1 - 1$ or $x_1 + 1$. We may as well guess that it's at $x_1 - 1$. 50% of the time we'd be right. Multiplying the two 50% we've got so far, that gives us a 25% chance so far of the confidence interval covering the right answer. The rest comes from the second line. If the surgeon manages to feel both toy social epistemologists then bingo! The toy Dean is definitely in the middle. But there's only a 50% chance of that happening. So altogether there's a 75% chance of the confidence interval covering the right answer. This means it's a valid 75% confidence interval, using the words "valid", "covering" and "confidence interval" the way they're meant to be used in Neyman–Pearson confidence interval theory.

Unfortunately, this 75% confidence interval is completely useless. By the time the surgeon's calculated the confidence interval, she knows for sure whether she's managed to feel one toy or two. (She must, or she wouldn't know the numbers to put into the confidence interval.) If she's only managed to feel one then she knows there's only a 50% chance that the confidence interval contains the position of the toy Dean, not a 75% chance. She should be looking for the missing toy on both sides of the one she's felt, not on the side which the confidence interval seems to recommend. And if she's managed to feel two toys then she knows exactly where the toy Dean is (right in the middle). In that case, to act on the basis of the 75% confidence interval (as if it only had a 75% chance of containing the position of the toy Dean, when actually it has a 100% chance) would be stupid and possibly fatal. Any reader who doesn't like complications should focus on the second possibility. *If the surgeon feels two toys, then she knows for sure where the third toy is, but confidence interval theory, because of the way it sets up its formulas in advance (which is an essential part of the theory), gives her only a 75% confidence interval for its position.*

Recall Langman's incorrect claim in the Lancet that "there is only a 2.5% chance that the true difference in the population at large" is to the right of a 95% confidence interval (Langman 1986, 716). If Langman's interpretation of confidence intervals had been right, there would be a 12.5% chance that the toy Dean was to the right of the 75% confidence interval we've calculated. But by the time the surgeon calculates her confidence interval she knows *for sure* that there is a 25% chance or a 0% chance (depending on whether she felt one toy or two) that the remaining toy is to the right of

the confidence interval. In the case we're considering, Langman's interpretation is not only wrong on theoretical grounds but also numerically way out.

My analysis of this example is not contentious (although, for some reason, my claim that it has implications for EBM is more contentious). According to one popular reading of the original theory of confidence intervals (a reading which I endorse, by the way, although the death of the author in both senses makes its truth irrelevant for present purposes), confidence intervals were never meant to tell us anything about the value of the parameter in a particular experiment. This is the reading of the theory which statisticians tend to fall back on when pressed to explain examples like this one. All a confidence interval can tell us, they say, is that in experiments similar to this one a certain result will happen 75% of the time, even though we may know *for sure* that it has not happened this time. A fat lot of use that is to the individual patient whom Sackett claimed to have in mind in the quote with which I opened the paper.

One more technical thought is necessary before we finish with the toy example. The problem can be avoided by calculating two separate intervals: one for the possibility that the surgeon feels one toy, and one for the possibility that she feels two toys. That solves the problem in this case, in an ad hoc way; can it be generalised? There are (surprisingly, not very many) more thoughtful advocates of confidence intervals who consider adapting confidence intervals in some such way; but it doesn't work, at least not in general. *Some* such counterexamples go away if one uses a technique known as conditioning on ancillary statistics, but many do not (Leslie 1998). If we avoid all of these sorts of counterexamples by only using confidence intervals when we find out, retrospectively, that they give sensible answers, the whole of confidence interval theory falls down; not only for the plausible reason that theories should not depend on such gerrymandering but also, demonstrably, because the theory essentially depends on not calculating its probabilities retrospectively (Berger and Wolpert 1988).

Confidence Intervals: Conclusions

It has been known to statisticians since at least the 1940s that confidence intervals *sometimes* give results which are known in advance to be completely wrong. I have to suppose that the few supporters of EBM who are aware of this problem take the attitude that only *sometimes* is not a problem. But there is an interesting catch here. There is no general theory which can tell us when a confidence interval will be subject to the sort of problem illustrated above ... except for very powerful theories (such as Bayesianism) which not only tell us that but also provide us with replacement estimates and significance tests. I do not have space here to go into these alternative theories in detail, but the details of Bayesianism are well known; work on other theories which might turn out to be equally powerful is in its infancy but an outline is given in (Barnett 1999). So it seems to me (for this and, in fact, for other more contentious reasons which I explore elsewhere) that we ought to be using these other theories to find out when confidence intervals will let us down ... and that we then might as well go the whole hog and use those theories as full replacements for confidence intervals. In other words, it seems to

me that any suspicion which the above example might instill that confidence intervals are not the best method of analysing experimental data is entirely correct.

My recommendation of somewhat vague alternative methods is necessarily a little wishy-washy, given the current state of statistical theory and of the philosophy of statistics. But that is not such a terrible thing, nor really such a surprising thing; and to have made a lack of clear-cut methodological rules seem surprising might be the worst thing EBM has done to us.

Methodological Pluralism

I have diagnosed two specific problems with the methodological doctrines of EBM. Perhaps more interestingly, I have diagnosed a problem with the fact that EBM *has* such specific doctrines, or rather that EBM involves applying such specific methodological doctrines so uniformly and over such an immense field.¹⁷

One possible antidote to the sort of situation we find ourselves in here is to let those whose tendencies run to methodological generalisation put together proposals like the worst excesses of EBM, and then stand on the sidelines and pick holes in them, as I have done in this paper so far. That can be fun, but it seems likely to be ineffective.

Another possible antidote is to propose a more positive stance which can pre-empt methodological overgeneralisations. An obvious antidote along such lines would be methodological pluralism of some sort: that is, some sort of understanding that there is no single best methodology.

In principle there is, presumably, a best methodology if not for a whole field then at least for each research situation. Or so most of us suppose, and so I will allow for the sake of argument. So in principle we should not need to be methodological pluralists.

However, this is not much more use than it would be to say that in principle there is, presumably, a design somewhere in Platonic heaven for a Panacea Machine which would choose perfect treatments for all diseases. What can we conclude from that? Nothing very useful. We don't have such a machine, and in the absence of one we are treatment pluralists.

What do I mean by saying that we are treatment pluralists? First of all, I mean that we believe we should allow different treatments in different situations. That is obvious and uncontentious. But in fact, we are treatment pluralists in a broader sense than that: we (and by "we" I mean all major organised health regulatory systems) allow different doctors to recommend different treatments in what is essentially the *same* situation.

Evidence-Based Medicine is something of a move against treatment pluralism, but it is not completely opposed to it. Even according to the staunchest EBMers, there are many health problems for which it is irredeemably contentious what the best treatment is, at least in the short term, either because nobody knows what the best treatment is or, slightly more controversially but not much, because a variety of people are giving reasons for preferring one treatment or another and there is no clear process for deciding between them. (Not even the staunchest EBMer thinks that there is *always* a clear process for deciding between competing medical arguments, even though EBM is an attempt to find a clear process whenever possible.)

So we (and now by “we” I mean not just regulators but surely all of us) are treatment pluralists in the fairly strong sense that we approve of a social epistemology in which a variety of treatments can be recommended by a variety of authorities even when addressing the same set of medical circumstances (“same” either in the type or the token sense; if you like, the same disease in the same patient).

This suggests what I think is the right sort of methodological pluralism. By analogy with treatment pluralism, we can use “methodological pluralism” to refer to approval of a social epistemology in which different methodologists can recommend different methodologies in what is essentially the same situation. I think that is right. But it would be useful to have a more easily applicable definition, for which I suggest:

methodological pluralism is the view that in any local context in which none of a variety of scientific methodologies is *clearly and uncontentionally right*, researchers should not be discouraged from using any methodology for which they can provide a good argument (using “argument” in the broadest sense).

Note that this is not currently the case: a researcher who wishes to use a methodology not approved by EBM has no social opportunity to present a good argument for doing so. It is certainly not possible, for example, to have a paper published in a medical journal in which one uses a substantially unorthodox statistical methodology to analyse, for example, a clinical trial. At least, so it is widely believed and, of course, the perception is itself part of the problem. In any case, no journal has ever published a major analysis of, for example, a large clinical trial which used a statistical methodology substantially different from p-values or confidence intervals (such as a Bayesian analysis).¹⁸

This statement of methodological pluralism mentions a judgement of a methodology as being either *clearly and uncontentionally right* for a specific situation or not. This raises the question of who should make such a judgement or, perhaps more interestingly, who should be allowed to enforce such a judgement in the areas in which methodological pluralism matters (editorial policy, funding policy, health service policy etc.). This is a tricky question which unfortunately I do not have space to discuss here, except to mention the obvious option, which is that the judgement as to whether a methodology is *clearly and uncontentionally right* could be made by the same social institutions which currently judge methodological issues in health and medicine (editors, funding bodies, governmental organisations and academic reviewers and consultants). In that case, it would be an obvious requirement of methodological pluralism that the people holding the relevant roles in those institutions, and the mechanisms which those institutions instantiate, do their best to take the arguments in favour of a variety of methodologies at face value, rather than (as they presently do, insofar as EBM is now the norm) privileging one set of methodologies once and for all.

Under this proposal, some improvement in the direction of methodological pluralism would be fairly easy. Slightly unorthodox positions such as the anti-EBM positions I take in my discussions of the examples above are fairly easy to find in the literature if one looks.

However, *some* positions would inevitably be ignored; and to the extent that potentially good methodologies are ignored even by people who are trying to be

methodological pluralists, we would have to judge methodological pluralism to be a failure. A more thorough methodological pluralism could no doubt be achieved by some reform of the institutions and social structures within which methodological judgements are made, perhaps along mildly Feyerabendian lines; but a reasonable discussion of that option would take me too far afield.

I do not know how to defend methodological pluralism in principle. I only know how to defend it in practice. This is like saying that I do not know how to defend treatment pluralism in principle, because I would not be able to defend it if we actually had a Panacea Machine. Lacking a Panacea Machine, in the here and now, we are all treatment pluralists; and all I want to do is defend methodological pluralism for the here and now (and, admittedly, for the foreseeable future) in which we do not have a methodology which is clearly and uncontentionally right even in specific, well defined situations, never mind one which is clearly and uncontentionally right across the whole of the type of large category into which EBM likes to sort experimental situations. The bulk of this paper has been an illustration of two ways in which we fail to have any *clearly and uncontentionally right* methodology. In both examples, I have canvassed two methodological views: the view which I claim is held by most proponents of EBM, and my own. Neither view is *clearly and uncontentionally right*. I have shown that the EBM methodology is not *clearly and uncontentionally right*, because it is wrong. But my own view is not *clearly and uncontentionally right* either, not because it is wrong but because it is contentious. In such cases, methodological pluralism is self-evident: all it says is that any methodology which has a good argument going for it should get a fair hearing. In particular, EBM should get a fair hearing, if only on the grounds that anything so widely believed has some chance of being right, no matter how strong my arguments that it is wrong seem to be; and alternatives to EBM should get fair hearings, because of the criticisms of EBM mentioned above and elsewhere. So methodological pluralism is what we need.

Acknowledgements

I would like to thank Adam La Caze, Joan Leach, Fiona Mackenzie, Alison Moore, Dick Parker, an anonymous reviewer for *Social Epistemology*, and the participants in the University of Queensland's Fourth Biohumanities Workshop, for helpful comments on this paper. The University of Queensland's Fourth Biohumanities Workshop was generously supported by the University of Queensland's Biohumanities Program and the University of Sydney's Centre for Time.

Notes

- [1] One of the few major differences between the various evidence hierarchies is that some of them have split what they call level 1 evidence into sub-levels, one of which corresponds to "meta-analyses", i.e., statistical amalgamations of a number of experiments on a single issue. This is a good idea. It happens not to affect any of the topics discussed in this paper.
- [2] To find my RCT and non-RCT, I have gone back in time a few years. As EBM has taken hold, it has become increasingly hard to find non-RCTs in the peer-reviewed medical literature.

(This in itself tells a story about the success of EBM, of course.) There are still some, and there is some hope that the tide is currently turning (thanks to Fiona Mackenzie for reminding me of this); but in any case, by going way back, I have been able to more or less match my RCT and my non-RCT in subject matter, which aids the comparison.

- [3] I'm taking a few things about how school anti-drug campaigns work for granted here; for example, that school students talk to each other.
- [4] Sometimes a large number of small RCTs can be combined in a statistical meta-analysis to produce worthwhile results. That does not affect the conclusion of this section, which is merely that some RCTs are bad. Idiosyncratic RCTs, or RCTs which have been conducted sufficiently badly, will never become part of a meta-analysis.
- [5] I say "I bet" because I cannot actually test this without access to the raw data.
- [6] In any case, as far as I can tell from their paper, the only variables which Meese et al. tested were age and CD4 lymphocyte count, so no amount of statistical analysis could tell whether the intervention and control groups were matched on such important variables as native language. This shows the importance of measuring the right things, something which tends not to be mentioned in the hierarchies of evidence.
- [7] As I argued above for Meese et al., MacGowan et al.'s sample size may effectively have been the number of social units, not the number of people, for some purposes. In that case, their sample size was only seven. But even then, they did much better than the bad RCT, because unlike the RCT they were not relying on their sample size to balance variables between the intervention and control groups. If an RCT's sample size is too small because of a confusion between individual and social levels of analysis, all of its results are suspect, because the analysis of an RCT depends on variables being balanced in this way. But if MacGowan et al.'s sample size was too small for the same reason, no such assumption was made, and all that happened was that they may have confused the effectiveness of methadone clinics in general (based on the individual level of analysis, assuming a large sample size) with the correct conclusion, namely that they'd tested some methadone clinics which may have been idiosyncratically good (the social level of analysis, assuming a small sample size of seven clinics with a large sample size within each clinic). What was a fatal mistake in an RCT is a relatively minor mistake here.
- [8] The OCEBM hierarchy is hardly more recent than the US Preventive Services Taskforce hierarchy, but the latter is a very late (although otherwise typical) member of an earlier generation of hierarchies, as Figure 1 demonstrates. I did not use the OCEBM hierarchy in my analyses above because it is much more complicated and uses non-standard labelling of its categories. We will see that it tells essentially the same story in any case, except for one important improvement: it does take sample size into account, at least sometimes, by requiring narrow confidence intervals in some cases. See below for a discussion of confidence intervals.
- [9] "Systematic reviews" are statistical agglomerations of smaller studies: what I earlier called meta-analyses. There has always been a tendency in England to avoid the word "meta-analysis", which some English people find pretentious. For a while, the English tried to popularise the alternative term "overview"; that failed, so now they are trying "systematic review".
- [10] There is a matching category of economic analysis called "Absolute better-value or worse-value analyses" which would apply when, for example, an intervention made something cost more without incurring any benefit.
- [11] The OCEBM hierarchy is promoted by a web site which is focused on clinical health care, so the fact that the hierarchy is also to some extent focused on clinical health care is by no means a criticism in itself.
- [12] Incidentally, I do not believe that even the narrow claim is entirely correct, although clearly it is vastly more defensible than dumb claim 1 (so much so that I would not call the narrow claim dumb).
- [13] In criticising confidence intervals, I do not wish to imply that researchers should return to previous methods of significance testing. Most, and possibly all, of the problems with

confidence intervals apply *mutatis mutandis* to standard hypothesis tests, statistical significance, p-values etc., since (e.g.) p-values are strictly intertranslatable with (isomorphic to) the end-points of confidence intervals, with only a few exceptions. When these exceptions occur, there is great (and interesting) uncertainty in the medical community about which is right, the p-value or the confidence interval. See Grossman et al. (1994) for an example.

- [14] There are statistical procedures which *can* assign probabilities to hypotheses, although they have their own drawbacks. The best known are Bayesian procedures. These days Bayesian methods are usually criticised for being excessively subjectivist, but non-subjectivist Bayesian methods exist and, in fact, were one of the main foils against which confidence intervals were originally intended to compete (Neyman 1937).
- [15] This example has been discussed to death in the mathematical literature (Welch 1939; Robinson 1975, 1977, 1979; Berger and Wolpert 1988), so I can be sure that there are no lurking infinities or other gotchas.
- [16] I say “a 75% confidence interval” rather than “*the* 75% confidence interval” to avoid wasting space by discussing the fact that confidence intervals are not unique. Even though they are not unique, the reader can trust me that there is no other confidence interval which makes more sense of this example. Note by the way that there is *no* 95% confidence interval in this case, unless we define one which is even more artificial and problematic than the 75% confidence interval.
- [17] I do not claim that it is the fault of the progenitors of EBM that it has ended up covering such an immense field. To some extent, the colonies of EBM may have sucked the colonists onto themselves. This interesting and important question is not my current topic.
- [18] Personal communication, Mahesh K. B. Parmar, Cancer Division, Medical Research Council Clinical Trials Unit, 13 July 2007; Donald A. Berry, M. D. Anderson Cancer Center, University of Texas, 15 July 2007

References

- Anonymous. 1987. Report with confidence. *The Lancet* 329 (8531): 488.
- Armstrong, J. Scott. 2007. Significance tests harm progress in forecasting. *International Journal of Forecasting* 23: 321–7.
- Barnett, Vic. 1999. *Comparative statistical inference*, 3rd edition. New York: John Wiley.
- Berger, James O., and Robert L. Wolpert. 1988. *The likelihood principle*, 2nd edition. Hayward, CA: Institute of Mathematical Statistics.
- Berry, Geoffrey. 1986. Statistical significance and confidence intervals. *Medical Journal of Australia* 144: 618–9.
- Calsyn, Donald A., Andrew J. Saxon, George Freeman, and Stephen Whittaker. 1982. Ineffectiveness of AIDS education and HIV antibody testing in reducing high risk behaviors among injection drug users. *American Journal of Public Health* 82: 573–5.
- Centre for Evidence-Based Medicine. u.d. Levels of evidence [cited 1 June 2008]. Available from <http://www.cebm.net/index.aspx?o=1025>; INTERNET.
- Daly, Jeanne. 2005. *Evidence-based medicine and the search for a science of clinical care*. Berkeley and Los Angeles: University of California Press.
- Evidence-Based Medicine Working Group. 2002. *Users' guide to the medical literature: A manual for evidence-based clinical practice*. Chicago: AMA Press.
- Gardner, M. J., and D. G. Altman. 1986. Statistics in medicine: confidence intervals rather than p values: estimation rather than hypothesis testing. *British Medical Journal* 292: 746–50.
- GRADE working group. 2004. Grading quality of evidence and strength of recommendations. *British Medical Journal* 328: 1490–7.
- Grossman, Jason, and Fiona J. Mackenzie. 2005. The randomised controlled trial: Gold standard, or merely standard? *Perspectives in Biology and Medicine* 48 (4): 516–34.

- Grossman, Jason, M. K. B. Parmar, D. J. Spiegelhalter, and L. S. Freedman. 1994. A unified method for monitoring and analysing controlled trials. *Statistics in Medicine* 13: 1815–26.
- Landis, S. E., J. L. Earp, and G. G. Koch. 1992. Impact of HIV testing and counseling on subsequent behavior change. *AIDS Education and Prevention* 4: 61–70.
- Langman, M. J. S. 1986. Towards estimation and confidence intervals. *British Medical Journal* 292: 716.
- Leslie, Claire F. 1998. Lack of confidence: A study of the suppression of certain counter-examples to the Neyman-Pearson theory of statistical inference with particular reference to the theory of confidence intervals. Master's thesis, University of Melbourne.
- Liddle, J., M. Williamson, and L. Irwig. 1996. *Method for evaluating research and guideline evidence*. Sydney: NSW Department of Health.
- MacGowan, Robin J., Robert M. Brackbill, Deborah L. Rugg, Nancy M. Swanson, Beth Weinstein, Alfred Couchon, John Scibak, Susan Molde, Paul McLaughlin, Thomas Barker, and Rich Voigt. 1997. Sex, drugs and HIV counseling and testing: A prospective study of behavior-change among methadone-maintenance clients in New England. *AIDS* 11: 229–35.
- Medicare Services Advisory Committee. 2000. *Funding for new medical technologies and procedures: Application and assessment guidelines*. Canberra: AusInfo.
- Meese, Peter, Olga Vujovic and Christopher Fairley. 1997. Randomised trial of a patient held record on the communication with general practitioners in patients with HIV infection. *Venereology* 10 (4): 226–7.
- National Health and Medical Research Council. 1998. Report of the 129th Session Sydney, 16 November 1998 [cited 1 June 2008]. Available from <http://www.nhmrc.gov.au/publications/reports/129sess.htm>; INTERNET.
- Neyman, Jerzy. 1937. Outline of a theory of statistical estimation based on the classical theory of probability. *Philosophical Transactions of the Royal Society of London, Series A* 236 (767): 333–80.
- Preventive Services Task Force. 1996. *Guide to clinical preventive services: Report of the US Preventive Services Task Force*. Baltimore: Williams and Wilkins.
- Robinson, G. K. 1975. Some counterexamples to the theory of confidence intervals. *Biometrika* 62 (1): 155–61.
- Robinson, G. K. 1977. Corrections and amendments: Some counterexamples to the theory of confidence intervals. *Biometrika* 64 (3): 655.
- Robinson, G. K. 1979. Conditional properties of statistical procedures. *The Annals of Statistics* 7 (4): 742–55.
- Sackett, David L., William M.C. Rosenberg, J.A. Muir Gray, R. Brian Haynes, and W. Scott Richardson. 1996. Evidence based medicine: What it is and what it isn't. *British Medical Journal* 312: 71–2.
- Savage, L. J., and discussants. 1962. *The foundations of statistical inference*. London: Methuen.
- Stuart, Alan, J. Keith Ord, and Steven Arnold. 1999. *Kendall's advanced theory of statistics vol. 2A: Classical inference and the linear model*, 6th edition. London: Arnold.
- Welch, B. L. 1939. On confidence limits and sufficiency, with particular reference to parameters of location. *Annals of Mathematical Statistics* 10: 58–69.