 DOI: http://dx.doi.org/10.5334/opt.111103

DO RCTS PROVIDE BETTER EVIDENCE THAN OBSERVATIONAL STUDIES?

By Talha Sami & Philip Sedgwick

Do not consider it proof just because it is written in books, for a liar who will deceive with his tongue will not hesitate to do the same with his pen.
Maimonides

Evidence Based Medicine (EBM) was born on the premise that some commonly used treatments are not evidence-based. The EBM philosophy is that Randomised Clinical Trials (RCTs) alongside systematic reviews offer stronger evidential support than observational studies, mechanistic reasoning and expert judgement (Howick, 2011, 10).

This paper argues that RCTs provide better evidential support than observational studies in two halves. Section I outlines EBM principles: it dissects RCTs and observational studies from an internal perspective. The foundation of Section I is that as long as RCTs entail less confounding factors than observational studies, they will provide superior evidential support[1] (Howick, 2011, p.61).

Section II will critique EBM ideology from an external perspective: if RCTs survive the criticisms then it is established that RCTs provide better evidential support than observational trials. It will start by analysing general principles and progress into scrutinising specific RCT devices. The conclusion is that RCTs withstand the criticism and are superior.

Section one – internal analysis

Before delving into any deeper theory, it is already apparent there is an intrinsic flaw within EBM ideology: it condemns expert opinion to the lowest form of evidence but the EBM hierarchy is *based* on expert opinion. This is a profound weakness and it cannot be fully answered; the reason this is so fundamental is because if the theory is questionable then surely so are the tenets of theory. This has been acknowledged by the proponents of EBM (Howick, 2011, 12) and that must be credited. However Howick, always so keen to defend EBM, ignores this weakness. The only neutral opinion is to judge the deeper tenets of EBM philosophy by their own merit.

Defining randomised controlled trials and observational studies

Firstly it is important to understand the purpose of a trial: classical statistical theory dictates that it is to reject the null hypothesis[2]. If the null hypothesis can be rejected in a traditional superiority

[1] The benefits of double masking and placebo deserve to be explored in their own right and this paper is not the forum for that however, if proven then the argument for RCTs would be doubly or triply stronger.

[2] The null hypothesis typically states that there is no difference between the treatments in question.

trial[3] then there is statistical significance between treatments (Ashcroft, 1999, 220). Let us consider the differences between observational trials and RCTs.

Observational studies are case-studies, cohort studies or historically controlled studies – they often involve looking at hospital records. They involve no intervention and the researchers observe natural variation. The assessors observe lifestyle choices and the impact these choices have on health; that is the impact of self-inflicted interventions. An example could be alcohol intake. One weakness of this type of study is that one cannot be sure if the observed difference in groups is down to the actual lifestyle choice or other differences between groups in their composition of variables including age, sex or disease severity.

Experimental trials include RCTs and crossover designs. The former are the focus in this paper – here an experimental intervention is compared with a control intervention that is a 'placebo'[4] or no treatment. In RCTs, all the participants are randomised at baseline so the group composition differences are minimised; this is a potential strength over observational trials, as the latter do not randomise. Multiple trials can be collated to form meta-analyses; the zenith of the EBM hierarchy. Systematic reviews can also inform the EBM hierarchy, but they are not as evidentially strong as meta-analyses as they will not be based on trials that incorporate randomisation.

Confounding factors provide a potential alternative explanation for the result achieved (Howick, 2011, 44). A confounding factor has three features: (1) The factor affects the outcome – an example of this may be age, gender and so forth (2) The factor is unequally distributed between the experimental and control group (3) The factor is not part of the experimental intervention.

Howick (2011, 49) proposes that observational studies suffer from more confounding factors than RCTs: these are self-selection bias[5], allocation bias[6], and performance bias[7]. RCTs differ from observational studies because their randomisation eliminates allocation bias whilst the double masking[8] and 'placebo' controls minimise performance bias. Even Worrall, a chief critic of EBM, concedes that due to the intrinsic nature of RCTs they *at least* eliminate allocation bias (Howick, 2011, 58); this is something observational studies cannot do. Consequently RCTs are superior on this basis alone[9].

[3] Traditionally clinical trials have been performed as superiority trials: they attempt to establish if there is a difference between a new treatment and standard treatment or placebo. (Sedgwick, 2011)

[4] Although there is no solid definition for placebo, we use Howick's definition: A legitimate placebo control is one that contains all and only the characteristic features of the experimental therapy (2011, 99)

[5] This is when a particular patient is deliberately included or excluded from a trial.

[6] This occurs when specific patients are allocated to specific treatment arms for a particular reason.

[7] This bias occurs during the trial: it can be manifested through the patient or the caregiver and/or assessor. It would involve a deviation from what the trial intended.

[8] This is when neither the patient nor the caregiver know which intervention the patient is taking.

[9] Due to word constraints, it is difficult to adequately examine all three types of confounding factor. The benefits of double masking and placebo deserve to be explored in their own right and this paper is not the forum for that. However, if their strengths are proven then the argument for RCTs would be doubly or triply stronger.

Selection Bias

Worrall attacks this idea that RCTs entails less confounding factors. He looks at selection bias (2007, 1008) – certain patients can be handpicked to be included or excluded in the trial. This can be manifested through picking those who might most benefit from the treatment, or conversely, certain individuals may be excluded because the side effects might be too much. Either way this would overestimate the effectiveness of the treatment in question.

Firstly, Worrall must realise that observation studies are also prone to selection bias therefore his criticism is negated. Secondly, randomisation in RCTs helps minimise the effect of non-intervention factors on the end results – randomisation is not a feature of observational trials. Worrall's attempts to criticise EBM actually backfire and do not leave observational trials in a superior place.

Conclusion

The groundwork of Section I is that as long as RCTs entail less confounding factors than observational studies, they will provide superior evidential support (Howick, 2011, 61). It has been shown that selection bias is present in both types of study, whilst allocation bias is eliminated in RCTs: therefore RCTs have the upper hand.

Section two – a critique of EBM

Section II firstly analyses the overarching principles of EBM: it corrects Worrall's misapprehensions about EBM. Then it moves the target to more specific devices: these include systematic review, the external validity of RCTs, the value of randomisation and the Number Needed to Treat (NNT).

Current day EBM ideology

If the study was not randomised we'd suggest that your stop reading it and go on to the next article in your search
(Sackett et al, 1996, 108)

Worrall bases much of his criticism on what was once overconfident EBM ideology: he coins the phrase 'no RCT, no evidence' for this position (2007, 987). Worrall states that EBM believes RCTs are 'unbiased'. However, EBM's actual position is quite different from what Worrall has proposed - current-day EBM scholars would actually state that RCTs *minimise* bias, and their argument is that RCTs provide more evidence than observational studies, mechanistic reasoning and expert judgement. Worrall has constructed a misrepresentation of EBM's position in order to demolish it: this is known as a straw man fallacy. Worrall's analysis is not based on actual EBM theory so it is quite irrelevant. This theme resonates throughout his criticisms and this paper.

Let us examine current EBM ideology: the latest EBM work actually allows observational studies to be ranked higher then RCTs in certain circumstances. The Grading of Recommendations Assessment, Development and Evaluation (GRADE) system classifies studies based on evidence

(Howick, 2011, 40). Firstly studies are ranked a priori[10]: high if they RCT or low if they are observational trial. Secondly they can be upgraded[11] or downgraded[12]. Thirdly they are assigned an a posteriori level of evidence: high, moderate, low or very low. GRADE is a balanced, straightforward system - it has none of the haughtiness that EBM initially exhibited and that Worrall criticises. Such breakthrough developments are exemplary of the new EBM movement: as RCTs are ranked higher in the first step, they are more likely to end with a higher ranking signifying that they provide more evidential support than observational trials.

An analysis of systematic reviews

The EBM hierarchy classifies systematic reviews as the highest form of evidence because they are a collection of RCTs. If Worrall is able to expose some flaws here then he has a strong basis on which to launch further attacks, thus weakening the rest of the hierarchy.

He calls systematic reviews a 'dark art' (2007, 992) because they are full of 'complex protocols': he adds that these rules differ from account to account and the underlying rationale for classifying efficacy is 'unclear'. There may well be some truth behind this, but he offers no more depth – such vague criticisms are applicable to anything. The lack of detail disintegrates his argument leaving the hierarchy untouched.

Reductionism

Worrall continues to attack general EBM principles: he does this by drawing a comparison between medical theory and physics. He believes this to be a logical move because he feels physics is "undisputedly" the most successful science (2007, 989), therefore physics can be used as a benchmark. Ignoring the disputable assumption that physics is superior to all other sciences, it seems that Worrall is employing a form of reductionism: this is a philosophical concept where a large system is condensed into its smaller constituents. In this case it seems that Worrall wants to reduce medicine to have the same strong empirical basis as physics. This would essentially be considering medicine as nothing more than deep molecular phenomena of the body, pathology, medicines and so forth. The implication of this is that Worrall does not consider medicine holistically – this has some serious repercussions.

There are two fundamental criticisms to what Worrall is suggesting: firstly attempting to reduce medicine to molecular phenomena may make sense at some level, but where are the benefits of intricate explanations in an A&E situation where the patient is about to enter ventricular fibrillation? His criticism is not *at all* realistic. Sober (2000, 74-76) concurs that physics can provide the information in every case, but in each case it would either be near impossible to state the explanation, because of our ignorance or the time it would take. He notes (2000, 26) that just because biology is reducible, it does not necessarily mean that it is the best way to advance our

[10] A priori justification is made independent of observation and experience in contrast to a posterior justification, which does makes reference to observation and experience

[11] They can be upgraded if they show: large and consistent effects, dose-response gradient or if all plausible confounders would reduce the size of the effect

[12] They can be downgraded if they show: inconsistency, indirectness, imprecision or publication bias

understanding – how will biology be helped by thinking about quarks and space-time? Sober's opinion has found support with Okasha, who agrees that different scientific disciplines should be for explaining different types of phenomena (2002, 55).

Secondly, Worrall's criticism can also apply to observational studies. By reversing the argument, one can ask: how do observational trials have a role in physics? The first point shows how weak Worrall's criticism is and the second point completely destroys his argument: his criticisms have been cancelled, leaving RCTs still superior to observational studies.

External validity of RCTs

Mant (1999) and Worrall (2007) both attack the external validity of RCTs – if these trials cannot be valid to a population other than that in the trial, they have little benefit and observational studies are of greater use.

Mant (1999, 743) believes that individuals in RCTs are not representative of the general population. Worrall (2007, 995) also criticises the external validity by citing the example of benoxaprofen: an NSAID[13] that was tested on 18-65 year old population but prescribed to the elderly for musculoskeletal and arthritic pain resulted in many deaths form hepato-renal failure. Both Worrall and Mant feel that RCTs should only be internally valid otherwise generalisations can occur causing horrendous consequences.

Firstly, both Mant and Worrall's criticisms can be made of observational trials: they can suffer from selection bias and their results can be inappropriately generalised respectively. Specifically in response to Worrall, there is no logic in discrediting the trial when the error is on those who mistakenly applied results to inappropriate populations. Secondly, neither Mant nor Worrall have appreciated the larger picture that RCTs are intended for a specific target population and the application of their results should be as such. Their criticisms are cancelled out leaving us at square one.

Randomisation

In a randomised trial, the only difference between the two groups being compared is that of most interest: the intervention under investigation (Worrall, 2007, p. 993)
Mike Clarke, the Director of the Cochrane Centre, UK

Worrall moves his target from ideology to specific RCT devices – he critiques randomisation. He has two issues: firstly he states (2007, 1001) that EBM believes randomisation controls for all confounders – known and unknown. The defence from EBM practitioners would be that Worrall has used the straw man fallacy once again: Worrall has constructed a position that is not actually in line with current day EBM ideology, so his attacks are redundant. In fact, EBM proponents would argue that they feel that RCTs are not unbiased but *less* biased than observational trials. This would be enough to counter Worrall's criticism, however he anticipated this dismissal as a straw man (2007, 1006): he attempts to focus his attack on medical practitioners and not the philosophers in EBM. His reply is not a very good one because his

[13] Non-steroidal anti-inflammatory

attempt to hone his criticism is purely semantic and does not actually mean anything. In this case, the EBM field was well prepared for such a critique and defended their ground well.

The second issue (Worrall, 2007, 1004) is that it is ridiculous to assert that through randomisation all known and unknown factors will be balanced within the two groups, leaving only one factor to be tested. The fact that unknown factors are unknown means they cannot be matched for (Worrall, 2007, 1003). This is a reasonable comment. EBM defenders would hone their argument: those confounders that can plausibly affect the outcome would be controlled for (Howick, 2011, 45). Worrall might reply, who defines what is 'plausible'? This type of philosophical to-and-fro serves no purpose: ultimately those who conduct current medical trials know what is important and will control for it. The weakness lies in conditions that little are known about, because then there is no benchmark to know what to control for. This niggling criticism has to be accepted but it does not by any means deconstruct EBM – in fact this should push EBM academics onwards.

Despite Worrall's attempts to deconstruct RCTs, he admits that even in observational studies there can only be matching of controls for known confounders (2007, 1010). In addition, he does admit the benefit of randomisation is to rule out allocation bias (Howick, 2011, 58). Consequently the superiority of RCT is still maintained.

The number needed to treat

Other assessments of RCTs also focus on particular devices – the Number Needed to Treat (NNT). It is a commonly used tool to assess the efficacy of a treatment when discussing treatment options with patients. It is a fundamental feature of RCTs and not frequently associated with observational studies. Sedgwick (2011) defines NNT as: "The reciprocal of the absolute risk difference in the primary outcome between the intervention and control groups".

McAlister's analysis (2008, 6) of NNT highlights some interesting limitations: (1) It is not to be considered in isolation but it is ideally used when comparing two treatments (2) It is best applied to acute conditions without any long-term repercussions and not any chronic conditions (3) It can be affected by baseline risk[14], time frame[15] and outcomes[16] (4) It can only be used for binary outcomes and not qualitatively. In addition to these four shortcomings, he states NNT can lead to 'misleading' and 'erroneous conclusions' whilst at other times it can be 'difficult to understand'. Given that there are so many limitations and restrictions upon it, the question is

[14] NNT varies inversely with baseline risk so is rarely favourable if evaluated in low risk populations. NNT will be larger if co-interventions reduce frequency of the outcome.

[15]NNT depends on when outcomes are counted. If relative risk reduction from long-term therapy is constant over time the NNT will decrease with increasing follow up as events accrue and absolute event rate increases. As time goes on increasing contribution from competing risks and concurrent medications may impact too.

[16]Most therapies impact on more than one outcome, therefore more than 1 NNT needs to be incorporated into treatment.

how can it be used effectively in a clinical environment? It might seem beyond belief that he ultimately concludes in the favour of NNT: he suggests a reflection on the limitations to maximise the efficiency of implementation. The sceptic might call for an alternative measure in the light of these limitations.

McAlister spends the majority of his paper looking at the weaknesses. It must be commended that he honestly assesses NNT. The limitations delineate when it is best to use the NNT: it helps counsel patients about certain treatments and can be used to compare two or more therapies that have been tested on a similar population in similar conditions. NNT permits a pragmatic comparison between treatments if there is a substantial difference between treatments in dropout rates; this means it has a real implementation in medicine. McAlister is clearly aware of the limitations of NNT and plays to its functions therefore he believes NNT can be of genuine use. The fact remains that there is no current substitute[17]. Indeed there may be some weaknesses, but that does not change the fact that NNT is the best theory to implement when the circumstances arise.

[17] He acknowledges that proteomics and pharmacogenomics are potential future replacements for NNT (despite their own pitfalls of subgroup analyses), when treatments can be personalised to the patient.

Concluding Remarks

The fact that the EBM hierarchy allocates professional opinion to the lowest tier, but EBM ideology itself is based on professional opinion is a serious cause for concern. In spite of this, EBM has developed a strong set of arguments and defence: in this light it is worthwhile to dismiss the aforementioned philosophical contradiction and focus solely on the empirical aspects of EBM. Section I reasoned that RCTs are superior for many reasons – even Worrall accepts that at least allocation bias is eliminated in RCTs but not observational trials therefore on this alone RCTs provide better evidence.

The critique in Section II targeted general principles, all the way down to particular devices of RCT. Worrall's criticisms often cancel out or are applicable to observational trials, so EBM ideology generally withstands criticisms - maintaining RCT superiority. Worrall's criticisms appear dismissive of EBM, but in fact it seems he is only trying to show that there can be flaws within RCTs - something the new EBM would not dispute.

The initial EBM ideology entailed some flaws; many of those are targeted by the critics. The fact remains that the criticisms that were once made targeted an old EBM: this is most notable with Worrall's critique. Since then the ideology has been strengthened and developed, therefore much of the criticisms made are now irrelevant.

Science & Technology Studies

Bibliography

Ashcroft R., (2002) 'What is clinical effectiveness?' *Stud. Hist. Phil. Biol. & Biomed. Sci* (33): 219-233

Howick J., (2011*) Resolving the paradoxes in Evidence-Based Medicine. A philosophical inquiry*. London: Wiley-Blackwell, pp. 12

Howick J., (2011) *Resolving the paradoxes in Evidence-Based Medicine. A philosophical inquiry*. London: Wiley-Blackwell, pp. 40

Howick J., (2011) *Resolving the paradoxes in Evidence-Based Medicine. A philosophical inquiry*. London: Wiley-Blackwell, pp. 44

Howick J., (2011) *Resolving the paradoxes in Evidence-Based Medicine. A philosophical inquiry*. London: Wiley-Blackwell, pp. 45

Howick J., (2011) *Resolving the paradoxes in Evidence-Based Medicine. A philosophical inquiry*. London: Wiley-Blackwell, pp. 58

Howick J., (2011) *Resolving the paradoxes in Evidence-Based Medicine. A philosophical inquiry*. London: Wiley-Blackwell, pp. 61

Howick J., (2011) *Resolving the paradoxes in Evidence-Based Medicine. A philosophical inquiry*. London: Wiley-Blackwell, pp. 99

Mant D., (1999) 'Can randomised trials inform clinical decisions about individual patients?' *The Lancet* (353): 743

Mautner T., (2005). *Dictionary of Philosophy* (2nd edn). London: Penguin Books

McAlister FA., (2008) 'The number needed to treat turns 20 – and continues to be used and misused'. *CMAJ* (179): 6

Okasha S., (2002) *Philosophy of Science: A very short introduction*. Oxford: Oxford University Press, pp. 56-57

Sackett DW., Richardson W., Rosenberg W., Haynes RB., (1996), *Evidence Based Medicine*. London: Churchill-Livingstone, pp. 108

Sedgwick P., (2011) 'Number needed to treat II'. *BMJ:* 342: d2644

Sedgwick P., (2011) 'Superiority trials'. *BMJ:* 342: d2981

Sober E., (2000). *Philosophy of Science* (2nd edn). USA: Westview Press, pp. 26

Sober E., (2000). *Philosophy of Science* (2nd edn). USA: Westview Press, pp. 74 – 76
Worrall J., (2007). 'Evidence in Medicine and Evidence-Based Medicine'. *Philosophy Compass* (2/6): 987

Worrall J., (2007). 'Evidence in Medicine and Evidence-Based Medicine'. *Philosophy Compass* (2/6): 992

Worrall J., (2007). 'Evidence in Medicine and Evidence-Based Medicine'. *Philosophy Compass* (2/6): 993

Worrall J., (2007). 'Evidence in Medicine and Evidence-Based Medicine'. *Philosophy Compass* (2/6): 995

Worrall J., (2007). 'Evidence in Medicine and Evidence-Based Medicine'. *Philosophy Compass* (2/6): 989

Worrall J., (2007). 'Evidence in Medicine and Evidence-Based Medicine'. *Philosophy Compass* (2/6): 1006

Worrall J., (2007). 'Evidence in Medicine and Evidence-Based Medicine'. *Philosophy Compass* (2/6): 1008

Worrall J., (2007). 'Evidence in Medicine and Evidence-Based Medicine'. *Philosophy Compass* (2/6): 1010

Worrall J., (2007). 'Evidence in Medicine and Evidence-Based Medicine'. *Philosophy Compass* (2/6): 1018