The evidence that evidence-based medicine omits

Brendan Clarke, Donald Gillies*, Phyllis Illari*, Federica Russo**, Jon Williamson****

**Department of Science and Technology Studies, University College London*
***Center Leo Apostel, Vrije Universiteit Brussel & Centre for Reasoning, University of Kent*
****Department of Philosophy, University of Kent*

Abstract: According to current hierarchies of evidence for EBM, evidence of correlation (e.g., from RCTs) is always more important than evidence of mechanisms when evaluating and establishing causal claims. We argue that evidence of mechanisms needs to be treated alongside evidence of correlation. This is for three reasons. First, correlation is always a fallible indicator of causation, subject in particular to the problem of confounding; evidence of mechanisms can in some cases be more important than evidence of correlation when assessing a causal claim. Second, evidence of mechanisms is often required in order to obtain evidence of correlation (for example, in order to set up and evaluate RCTs). Third, evidence of mechanisms is often required in order to generalise and apply causal claims.

While the EBM movement has been enormously successful in making explicit and critically examining one aspect of our evidential practice, i.e., evidence of correlation, we wish to extend this line of work to make explicit and critically examine a second aspect of our evidential practices: evidence of mechanisms.

All studies are fallible

The EBM movement views evidence of mechanisms as poor quality evidence. (Terminology: '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). 'A mechanism for a phenomenon consists of entities and activities organized in such a way that they are responsible for the phenomenon' (Illari and Williamson 2012, p. 120). Here evidence of mechanisms may include evidence obtained by laboratory studies or previous statistical studies and tends to be relayed by expert testimony, e.g., via scientific publications.) This dim view of mechanistic evidence is most obvious when one refers to the 2011 Levels of Evidence table issued by the Oxford Centre for Evidence Based Medicine (OCEBM 2011), which places 'mechanism-based reasoning' at level 5 – the lowest level – of the hierarchy of evidence. (Here, 'Mechanistic reasoning is an inferential chain (or web) linking the intervention (such as HRT) with a patient-relevant outcome, via relevant mechanisms' (Howick 2011, p.929).) Earlier evidence hierarchies, although often less explicit, also tend to leave only one possible place for prior evidence of mechanisms: the bottom level. For Canadian Task Force (1979, p.1195), for instance, levels I and II are occupied by statistical studies and everything else is relegated to the bottom level: 'III: Opinions of respected authorities, based on clinical experience, descriptive studies or reports of expert committees'.

Higher up these hierarchies of evidence come various kinds of statistical study, including controlled studies, randomised controlled studies and, at the apex, systematic reviews and meta-analyses. The thought is that for these statistical studies, the higher up the hierarchy the better the evidence copes with the problem of confounding. (This is the problem that an observed dependence between *A* and *B* may be attributable to variation in *B*'s other causes, rather than variation in *A*.) But it is

important to note that no statistical study *solves* the problem of confounding. A randomised controlled experiment only yields treatment / non-treatment groups that are homogeneous with respect to the putative effect's other causes in the asymptotic limit, as the number of individuals assigned to each of these two groups goes to infinity. So, while the evidence hierarchies may correctly identify the *relative* merits of various kinds of statistical study for dealing with the problem of confounding, in *absolute* terms these studies are all very fallible. There is room, therefore, for other kinds of evidence to influence decision making, even when high quality RCT evidence is available.

In particular, since statistical studies are fallible, strong evidence of mechanisms can sometimes override evidence gleaned by a statistical study that is high up in the hierarchy. This is especially clear when there is strong prior evidence that there is *no* mechanism linking the putative cause with the putative effect. In this case, the best remaining explanation is that the increase in the effect variable is due to confounding. Thus it can be reasonable to dismiss the claim that remote, retroactive intercessory prayer shortens length of stay in hospital, despite evidence from an RCT that yields a significant correlation between the two variables (Leibovici 2001), on the grounds that current science holds no place for any mechanism that can explain the putative effect in terms of the putative cause. Similarly for claims made in favour of precognition on the basis of a report of 9 experiments (Bem 2011) – positive results which eventually turned out not to be replicable (Richie et al 2012). Certain claims in favour of homeopathy – including positive results from systematic review and meta-analysis (Cucherat et al 2000) – can be treated analogously. While these examples are extreme, they clearly show that mechanistic evidence should not be confined to the bottom level of evidence hierarchies, but should, in certain cases, be considered alongside high-level statistical evidence.

Assessing RCTs

Let us now consider some simple historical reasons for our position, before returning to our philosophical argument in later sections. Historical examples suggest that the use of RCTs does not allow us to dispense with evidence of mechanisms, because evidence of mechanisms is needed to design and interpret RCTs. This is well illustrated by one of the first and most famous RCTs. This trial, carried out in the UK beginning in September 1947, was used to test whether streptomycin is an effective cure for pulmonary tuberculosis. The results of this RCT after 6 months were that, in the Streptomycin group of 55, 28 (51%) had shown considerable improvement and only 4 (7%) had died, whereas, in the control group of 52, only 4 (8%) had shown considerable improvement whereas 14 (27%) had died (MRC, 1948, p. 771). Seemingly the RCT was an overwhelming success, but there was also some evidence that the bacilli responsible for the disease were developing resistance to streptomycin. This led the scientists conducting the trial to express caution, and to recommend that the observation of patients involved in the trial should be continued. This caution, based on evidence relating to the mechanism of the disease, proved to be amply justified. After 5 years, 32 of the 55 in the streptomycin group had died (58%), compared with 35 of the 52 in the control group (67%) (Florey, 1961, p. 133). The difference here is not statistically significant, and this showed that, over a longer period, streptomycin, on its own, was no better than existing treatments. If evidence of mechanisms had not been taken into account, the misleading impression that streptomycin on its own was an effective therapy would have been given, and this would have delayed the development

of the first genuinely effective treatment, which was a combination of streptomycin with para-amino-salicylic acid (PAS).

As the streptomycin case demonstrates, evidence of mechanisms informs the design and interpretation of RCTs. While it is theoretically possible to conduct an RCT in the absence of evidence of mechanisms – as in the case of Leibovici et al (2001) – most clinical trials do evaluate interventions that are somewhat mechanistically understood. Likewise, the manner in which this evaluation is performed – including the decision to clinically evaluate particular interventions; the way in which these interventions are carried out; and the measurement of the effects of these interventions. This makes evidence of mechanisms central to the business of conducting clinical trials. Note that this consideration of evidence of mechanisms does not mean that judgements of efficacy proceed on entirely mechanistic grounds (see e.g. Howick 2011: 128).

Given that clinical trials typically seek to investigate novel interventions, the evidence of mechanisms upon which interventions and outcome measures rely is highly dynamic and rapidly changing. This can be demonstrated by estimating the age of such measures in contemporary clinical trials. Of the ten most-cited articles from the last five years of *The Lancet* (Scopus data April 21 2012), the age at publication of both interventions and outcome measures used was estimated. Of these ten articles, collectively cited 6132 times, three (Black et al 2008; Daemen et al 2007; Goldberg et al 2008) were identified as non-RCT publications, and excluded. From the remaining seven, two each dealing with HIV-AIDS (Gray et al 2007; Bailey et al 2007) and renal cell carcinoma (Escudier et al 2007; Motzer et al 2008), and one each on HPV vaccination (Paavonen et al 2007), vascular outcomes in diabetes (Patel et al 2007), and breast cancer (Smith et al 2007), a total of 35 intervention or outcome measures were identified (see table 1, supplementary material). The age was estimated as per the methods discussed in the supplementary material. Where there was doubt about the introduction of a particular measure, the oldest recorded instance was used. The average age at the time of publication is 15 years, excluding those interventions or outcome measures thought to be older than 100 years, with the youngest intervention ranging between 1 and 23 years (mean=10).

The sheer novelty of these critical parts of trial construction indicate that, far from being background or common knowledge, the evidence used to build and interpret trials changes rapidly. Given that a central principle of EBM practice is the "...conscientious, explicit, and judicious use of current best evidence..." (Sackett et al 1996), we suggest that evidence of mechanisms should therefore be subject to the same process of systematic critical appraisal as evidence gleaned from trials themselves.

External validity

Even if we grant the soundness of an RCT, a question remains about its external validity. There is no a priori reason why the results of an RCT should be straightforwardly applicable to another population. This concerns medical treatments as well as policy actions. This problem is thoroughly discussed by Victora et al. (2004), where the authors point to several issues that hinder the external validity of RCTs. In particular, the authors dispute that the internal validity of an RCT also ensures its generalisability. The assumption that it does follows, Victora et al explain, from the assumption of

'universal biological response'. Victora et al. (2004) challenge this view and argue that although this assumption might well be hold for "interventions with short causal pathways", it is certainly not the case for "interventions involving long, complex causal pathways, or in large-scale evaluations where these pathways can be affected by numerous characteristics of the population, health system, or environment", such as policy interventions. In fact, there might be two threats to successful extrapolation in the case of policy: one is "behavioural effect modification" and the other is "biological effect modification" (i.e., respectively, "differences in the actual dose of the intervention delivered to the target population" and "differences in the dose-response relationship between the intervention and the impact indicator").

Cartwright (2011) makes a similar point and illustrates it with the example of the 'Bangladesh Integrated Nutrition Policy (BINP), a programme that largely failed to have an impact on child nutrition, although a very similar programme proved highly successful in Tamil Nadu (TINP – the Indian Tamil Nadu Integration Project). Cartwright makes the point that policy makers neglected the different social structure of the populations to which they applied the programme, and this explains the success in one case and the failure in another case. Social structures can in fact be understood in mechanistic terms too (see e.g. Demeulenaere 2011). Evidence of mechanisms helps assess the external validity of an RCT (or indeed of any study) because it adds precious knowledge about the similarities between the test and target populations. This point has been forcefully argued for by Steel (2008). 'Mechanism-based' external validity inferences are a significant step forward with respect to the Cook & Campbell tradition (Cook and Campbell 1979) that connects validity merely to the representativeness of the sample and to the possibility of replicating the study.

The problem of inferring from the population to the single case

There is also another sense in which external validity poses a problem. Above, we discussed the inference from one population to *another* population. Here, the issue concerns the inference from the population (studied in the RCT) to a *particular patient*. While it is a merit of the evidence-based movement to have fostered protocols for treatment in order to ensure standardisation and comparability, there is no a priori guarantee that an individual patient will be similar enough to the average individual of the RCT and that, consequently, s/he will respond to the treatment in the same way. In such cases, considerations to do with single-case individual responses will be vital to support a claim that the same treatment will work in the single case.

This kind of consideration is particularly vital in treatments for diseases where a variety of distinct causal mechanisms produce clinically similar effects. In the case of breast cancer, tumours may be distinguished by the kinds of receptors they express, and this classification is predicated on the different mechanisms at work in these tumours. Similarly, melanoma classifications now often include consideration of particular genetic mutations (Clarke 2011). Both these cases are motivated by therapeutic considerations: statistical evidence suggests that differently constituted tumours respond very differently to particular treatments. Thus one needs to know which mechanisms, or features of mechanisms, are instantiated in the particular patient. Again, statistical evidence works better to "make decisions about the care of individual patients"(Sackett et al 1996) when integrated with evidence of mechanisms.

Integration of evidence

What we really need is to use the totality of evidence available to us. When we must use fallible sources of evidence – and all sources of scientific evidence are fallible – it is better to look for independent converging sources of evidence, as a single good source of evidence will fail significantly often (Wimsatt, 2007). Evidence of correlations obtained from RCTs or observational studies and evidence of mechanisms are independent sources of evidence that are usefully complementary. We have shown that evidence of mechanisms supplements evidence of correlation in designing and assessing RCTs, and in inferring from population to population, and from a population to the single-case. A serious problem with evidence of correlation is the problem of confounding: e.g., when a correlation between variables *A* and *B* may be the result of a common cause of *A* and *B*. Tracing a mechanism from *A* to *B* helps alleviate that worry by offering a direct connection to account for the correlation.

The parallel problem is that evidence of a mechanism does not on its own establish an average causal effect between *A* and *B*. Evidence of one mechanism linking *A* and *B* cannot establish that there aren't other mechanisms linking *A* and *B*, which may balance out, or *mask*, the effect of the known mechanism. But evidence of correlation between *A* and *B* is exactly what is needed to address this masking problem. The best evidence that *A* causes *B* is evidence of a mechanism linking *A* and *B*, where the expected effect size between *A* and *B* is commensurate with the effect size observed in RCTs (if possible) or observational studies seeking a correlation between *A* and *B*. Evidence of mechanisms and evidence of correlation are complementary: each addresses the primary weakness of the other. What we advocate is a pragmatic evidential pluralism, which uses the totality of available evidence.

The problem that we have identified is not that mechanistic evidence is being ignored. Mechanistic evidence *is* being used to eliminate confounding, to set-up and interpret RCTs, and to extrapolate from one population to another. It is clear from informal discussions with researchers and those charged with approving drugs that mechanistic evidence is being used – often tacitly – alongside statistical evidence in order to establish causal claims. But all this happens *despite* the protocols offered by evidence hierarchies, which urge that, when good statistical evidence is available, it should be considered to the exclusion of other forms of evidence. Evidence hierarchies need revising to ensure that complementary forms of evidence are treated as complementary, and that evidence of mechanisms, currently treated implicitly, is examined explicitly.

References

Alcena V, 1986. AIDS in third world countries. N Y State J Med 86:446.

Bailey RC, Moses S, Parker CB, Agot K, Maclean I, Krieger JN, Williams CF, Campbell RT, Ndinya-Achola JO, 2007. Male Circumcision for HIV Prevention in Young Men in Kisumu, Kenya: a Randomised Controlled Trial. Lancet 369:643–656.

Bem DJ, 2011. Feeling the Future: Experimental Evidence for Anomalous Retroactive Influences on Cognition and Affect. J Pers Soc Psychol 100:407-425.

Black RE, Allen LH, Bhutta ZA, Caulfield LE, de Onis M, Ezzati M, Mathers C, Rivera J, 2008. Maternal and Child Undernutrition: Global and Regional Exposures and Health Consequences. Lancet 371:243–260.

Canadian Task Force on the Periodic Health Examination 1979. The periodic health examination. Can Med Assoc J 121:1193-1254.

Cartwright ND, 2011. Evidence, external validity and explanatory relevance, in: Morgan, G (Ed), The philosophy of science matters: the philosophy of Peter Achinstein, Oxford University Press, New York: 15-28.

Clarke B, 2011. Causation and melanoma classification. Theor Med Bioeth 32:19-32.

Cook T, Campbell D, 1979. Quasi-experimentation: Design and analysis issues for field settings, Rand MacNally, Chicago.

Cucherat M, Haugh MC, Gooch M, Boissel J-P, for the HMRAG group, 2000. Evidence of clinical efficacy of homeopathy: A meta-analysis of clinical trials. Eur J Clin Pharmacol 56:27-33.

Daemen J, Wenaweser P, Tsuchida K, Abrecht L, Vaina S, Morger C, Kukreja N, Jüni P, Sianos G, et al., 2007. Early and late coronary stent thrombosis of sirolimus-eluting and paclitaxel-eluting stents in routine clinical practice: data from a large two-institutional cohort study. Lancet 369:667–678.

Demeulenaere P (Ed.), 2011. Analytical sociology and social mechanisms, Cambridge University Press, Cambridge.

Escudier B, Pluzanska A, Koralewski P, Ravaud A, Bracarda S, Szczylik C, Chevreau C, Filipek M, Melichar B et al., 2007. Bevacizumab Plus Interferon Alfa-2a for treatment of metastatic renal cell carcinoma: a randomised, double-blind phase III trial. Lancet 370:2103–2111.

Fink A, 1986. A possible explanation for heterosexual male infection with AIDS. N Engl J Med 314:1167.

Florey ME, 1961. The clinical application of antibiotics. Volume II streptomycin and other antibiotics active against tuberculosis, Oxford University Press, London.

Goldenberg RL, Culhane JF, Iams JD, Romero R, 2008. Epidemiology and causes of preterm birth. Lancet 371:75–84.

Gordon MS, Margolin K, Talpaz M, Sledge GW Jr, Holmgren E, Benjamin R, Stalter S, Shak S, Adelman D, 2001. Phase I safety and pharmacokinetic study of recombinant human anti-vascular endothelial growth factor in patients with advanced cancer. J Clin Oncol 19:843-850.

Gray RH, Kigozi G, Serwadda D, Makumbi F, Watya S, Nalugoda F, Kiwanuka N, Moulton LH, Chaudhary MA, et al., 2007. Male circumcision for HIV prevention in men in Rakai, Uganda: a randomised trial. Lancet 369:657–666.

Howick J, 2011. Exposing the vanities—and a qualified defense—of mechanistic reasoning in health care decision making. Philos Sci 78:926-940.

Illari PM, Williamson J, 2012. What is a mechanism? Thinking about mechanisms across the sciences. Eur J Philos Sci 2:119-135.

Krieger JN, Bailey RC, Opeya J, Ayieko B, Opiyo F, Agot K, Parker C, Ndinya-Achola JO, Magoha GA, et al., 2005. Adult male circumcision: results of a standardized procedure in Kisumu District, Kenya. BJU Int 96:9–13.

Leibovici L, 2001. Effects of remote, retroactive intercessory prayer on outcomes in patients with bloodstream infection: randomised controlled trial. BMJ 323:1450-1451.

Motzer RJ, Escudier B, Oudard S, Hutson TE, Porta C, Bracarda S, Grünwald V, Thompson JA, Figlin RA, et al., 2008. Efficacy of everolimus in advanced renal cell carcinoma: a double-blind, randomised, placebo-controlled phase III trial. Lancet 372:449–456.

MRC, 1948. Streptomycin treatment of pulmonary tuberculosis. BMJ 2:769-782.

OCEBM Levels of Evidence Working Group. "The Oxford 2011 Levels of Evidence". Oxford Centre for Evidence-Based Medicine. Last access 24-09-2012 *http://www.cebm.net/index.aspx?o=5653.*

Paavonen J, Jenkins D, Bosch FX, Naud P, Salmerón J, Wheeler CM, Chow SN, Apter DL, Kitchener HC, 2007. Efficacy of a prophylactic adjuvanted bivalent L1 virus-like-particle vaccine against infection with human papillomavirus types 16 and 18 in young women: an interim analysis of a phase III double-blind, randomised controlled trial. Lancet 369:2161–2170.

Patel A, 2007. Effects of a fixed combination of perindopril and indapamide on macrovascular and microvascular outcomes in patients with type 2 diabetes mellitus (the ADVANCE Trial): a randomised controlled trial. Lancet 370:829–840.

Ritchie SJ, Wiseman R, French CC, 2012. Failing the future: three unsuccessful attempts to replicate Bem's 'retroactive facilitation of recall' effect. PLoS ONE 7: e33423. doi:10.1371/journal.pone.0033423

Sackett DL, Rosenberg WM, Gray JA, Haynes RB and Richardson WS, 1996. Evidence based medicine: what it is and what it isn't. BMJ 312:71–72.

Smith I, Procter M, Gelber RD, Guillaume S, Feyereislova A, Dowsett M, Goldhirsch A, Untch M, Mariani G, 2007. 2-year follow-up of trastuzumab after adjuvant chemotherapy in HER2-positive breast cancer: a randomised controlled trial. Lancet 369:29–36.

Steel D, 2008. Across the boundaries. Extrapolation in biology and social science, Oxford University Press, Oxford.

Victora CG, Habicht JP, Bryce J, 2004. Evidence-based public health: Moving beyond randomized trials. Am J Public Health 94:400–405.

Wimsatt WC, 2007. Re-engineering philosophy for limited beings: piecewise approximations to reality, Harvard University Press, Harvard.

Supplementary material

Name	Introduced	Intervention / outcome	Age in years at publication	Publication
Placebo	Pre-1900	I	>100	Escudier et al 2007
Placebo	Pre-1900	I	>100	Patel et al 2007
Placebo	Pre-1900	I	>100	Motzer et al 2008
Indapamide	1975	I	32	Patel et al 2007
Perindopril	1984	I	23	Patel et al 2007
Interferon alfa-2a	1985	I	22	Escudier et al 2007
Circumcision	1986	I	21	Gray et al 2007
Delayed circumcision	1986	I	21	Gray et al 2007
Trastuzumab	1989	I	18	Smith et al 2007
Havrix	1992	I	15	Paavonen et al 2007
Everolimus	1997	I	11	Motzer et al 2008
Bevacizumab	2001	I	6s	Escudier et al 2007
HPV16/18 L1 vaccine	2004	I	3	Paavonen et al 2007
Circumcision	2005	I	2	Bailey et al 2007
Delayed circumcision	2005	I	2	Bailey et al 2007
Progression-free survival	Pre-1900	O	>100	Escudier et al 2007
New or worsening diabetic eye disease	Pre-1900	O	>100	Patel et al 2007
Disease-free survival in HER2+ breast cancer	Pre-1900	O	>100	Smith et al 2007
CVD death	Pre-1900	O	>100	Patel et al 2007
Non-fatal stroke	Pre-1900	O	>100	Patel et al 2007
Non-fatal MI	Pre-1900	O	>100	Patel et al 2007
Progression-free survival	Pre-1900	O	>100	Motzer et al 2008
New or worsening renal disease	1936	O	71	Patel et al 2007
HIV immunoassay (Welcozyme)	1980	O	27	Gray et al 2007
HIV immunoassay (Vironostika)	1987	O	20	Gray et al 2007
HIV PCR	1988	O	19	Bailey et al 2007
HIV LIA	1991	O	16	Bailey et al 2007

HIV Western blot	1991	O	16	Gray et al 2007
HIV synthetic peptide test Determine HIV 1/2	1998	O	9	Bailey et al 2007
HIV rtPCT	1999	O	8	Gray et al 2007
HIV Unigold Recombigen HIV	2003	O	4	Bailey et al 2007
Cervical cytology with testing for 14 oncogenic HPV types by PCR Hybrid Capture 2	2003	O	4	Paavonen et al 2007
Colposcopic biopsy for 14 oncogenic HPV by PCR Hybrid Capture 2	2003	O	4	Paavonen et al 2007
HPV16 and 18 ELISA	2004	O	3	Paavonen et al 2007
HIV ELISA Detect HIV 1/2, Adaltis Inc	2006	O	1	Bailey et al 2007

Measures were dated using the following scheme. Procedures were dated in a contextual manner. Where a precise technique was specified in print, then the publication date of the resource referred to as descriptive of that procedure was used. For example, Bailey (et al 2007) specified the use of a circumcision technique dating from 2005 (Krieger et al 2005). Where no technique was specified, as in the case of Gray (et al 2007), the date at which that particular procedure might usefully be employed in the specified context was employed. In this case, the date was given as 1986, being the first suggestion that circumcision might be a means of preventing HIV transmission (Alcena 1986; Fink 1986). Diagnostic tests were dated by FDA approval date where available, otherwise by their earliest mention in PubMed. Drugs, likewise, were dated by first mention in PubMed for their name or synonyms in the general context of their use in the analysed trial. Note however that this did not take account of the context of that practice as either research or clinical. So bevacizumab is dated as 2001 because of its first use in advanced solid tumours at that time (Gordon et al 2001). Combinations of agents in a trial were recorded once only in a publication. For instance, a trial comparing the efficacy of bevacizumab plus interferon alfa-2a against placebo plus interferon alfa-2a (Escudier et al 2007) led to the separate dating of bevacizumab, interferon alfa-2a and placebo. However, the use of placebo in multiple publications leads to multiple recordings in the final data. Measures without clear date of introduction were pessimistically coded into pre-history, even when they deal with relatively newly described diseases or when they presumably rely on new investigation techniques. For instance, breast cancer is ancient, but HER2 was identified during the 1980s. Disease-free survival in HER2+ breast cancer is coded as >100 years. Where there was doubt about the introduction of a particular measure, the oldest recorded instance was used.