J. of Integrative Research & Reflection, Vol. 02, May XX, 2019

Confronting the Obvious: An Epistemological Examination of the Evidence Informing Evidence-Based Medicine

Dinesh Moro

Introduction

Following its rise in popularity through the late 90s and early 2000s, modern medical practice has enthusiastically embraced the idea of evidence-based medicine, or EBM (Claridge and Fabian 2005, 548). It is now the most widely accepted medical model, and that popularity has come with its share of criticism (Greenhalgh et. al 2014, Haynes 2002, Cohen and Hersh 2004). With the exception of Maya Goldenberg in her astute work, *On Evidence and Evidence-Based Medicine: Lessons from the Philosophy of Science*, few have examined the fundamental epistemological processes of evidence creation in EBM. The goal of this essay is to offer insight to both clinicians and patients by building upon Goldenberg's work with concrete examples of some of EBM's epistemological issues. In doing so, I will propose that those practicing EBM should scrutinize the nature of knowledge informing their profession so as not to place excessive trust in clinical research findings. Additionally, healthcare practitioners should augment their critical understanding of EBM by discussing its guiding epistemic values and by promoting epistemic humility throughout medicine. I will begin by fully explaining the EBM paradigm before moving into critiques of the model's portrayed objectivity and use of induction, the gold standard afforded to randomized controlled trials, and the label of statistical significance. Finally, I will briefly discuss some of the aforementioned recommendations.

What is Evidence-Based Medicine?

Before a critique of EBM can be undertaken, it is important that the reader knows what exactly is meant by the term and which claims about knowledge it implies. EBM is currently accepted as the best-practice model for healthcare practitioners to structure patient treatment (Goldenberg 2005, 2621). It is a relatively new paradigm, having only gained popularity in the late twentieth century (Claridge and Fabian 2005, 548). In 1996, the term was formally defined by McMaster University physician David Sackett as "the conscientious, explicit, and judicious use of current best evidence in making decisions about the care of individual patients" (Sackett 1996, 71). Sackett's evidence-based medicine took influence from Archie Cochrane's popular 1972 work,*Effectiveness and Efficiency: Random Reflections on Health Service*, similarly placing vital importance on the use of randomized control trials, or RCTs, to assess the effectiveness of treatments (Claridge and Fabian 2005, 552).

In line with its roots, modern EBM emphasizes that healthcare professionals should con-

stantly be updating their knowledge through reading new articles relevant to their practice, with special considerations given to RCTs. Sackett rationalizes that RCTs, and especially systematic reviews of several RCTs, are the gold standard for judging a treatment effect since they are "much more likely to inform us and so much less likely to mislead us" (Sackett 1996, 72). The status afforded to RCTs means that they have been ranked at, or near, the top of nearly fifty hierarchies of evidence (Schünemann 2006, 3). Cohort studies, case-control studies, and case series find themselves lower on these hierarchies, supposedly providing progressively weaker evidence (Burns et al. 2011, 8).

Supporters of EBM believe that by using the scientific method to assess common clinical practices, the efficacy of those practices can be determined. Sharing those experimental results then allows for clinicians to learn from a wider pool of patient interventions, informing their practice in a way that is more rigorous than relying on personal experience (Timmermans and Mauck 2005, 20). EBM is also thought to reduce the number of irregularities in clinical procedures, replacing them with best-practice guidelines. Ultimately, EBM is seen as a promising way to inform both patients and clinicians by providing them with high-quality evidence, which they can then use to inform their decisions.

In practice, these high hopes for EBM may not be panning out. In 2001, a study was conducted to determine the critical appraisal skills and state of knowledge with regard to EBM in a sample of 286 family physicians from Ontario (Godwin and Seguin 2003, 4). 95% of respondents saw EBM as important to the practice of medicine. However, test scores on general EBM knowledge (including the interpretation of results and research methods) were around 50% with the average score being just 6.4 out of 12 (Godwin and Seguin 2003, 5). Interestingly, younger physicians, aged 25-35, scored higher with an average of 8.2 out of 12 while the oldest cohort, aged 56-65, scored an average of 4.4 out of 12 (Godwin and Seguin 2003, 6). The researchers characterized the results as not impressive, especially given that physicians in the younger cohort would have learnt about EBM as part of their curriculum (Godwin and Seguin 2003, 6). These findings suggest that clinicians need to develop their ability to critically evaluate research. Appropriately, that is precisely the goal of this article.

Philosophy of Science Problematizes EBM

One of the primary tenets of EBM is that consulting current best evidence will improve patient care. Incumbent in this belief is that not all knowledge can be called evidence. A common perception is that evidence, and scientific knowledge in general, is derived from an adherence to the scientific method and produces facts (Chalmers 2015, *xx*). At a very fundamental level, EBM shares this attitude founded on the positivist principle that "knowledge should be derived from the facts of experience" (Chalmers 2015, 3).

Today, philosophers of science understand that this positivist outlook is not a comprehensive account of scientific inquiry. The first problem is that our observations cannot lead us to objective fact, as they are always to some degree coloured by our past experiences or background knowledge (Chalmers 2015, 7). The truth of this statement can be demonstrated in any radiology department as specialists examining x-rays are able to expound a host of relevant and detailed medical information while a non-trained observer may struggle to determine through which side of the image light is meant to pass (Chalmers 2015, 8). Critically, it can be understood that whenever an observation is made in the pursuit of science, we are observing phenomena not as they exist naturally but through a certain subjective lens. The same holds for medical research: clinical findings cannot be objective in a conventional sense. Researchers must make decisions on what should count as a specific outcome (i.e. whether the change in a patient's condition should be considered an improvement) and what

statistical analyses will best illuminate trends in their data.

It is very difficult to find studies that investigate how clinicians view research and objectivity. However, critiques of medical education have suggested that clinicians are often enculturated into an objectivist perspective by encouraging them to take on the role of "detached observer" when interacting with patients (Wilson 2000, 206). Writing in medical journals rarely contains first-person language, showing that perhaps this detached observer mentality carries on into professional work (Williams 2010, 214). Thus, it may come as a surprise to some clinicians that despite supposedly strong study design, the practice of EBM is not entirely objective or free from personal bias. Clinicians may then be encouraged to re-evaluate the trust they place in research findings.

Questioning the Reputation of Randomized Controlled Trials (RCTs)

The RCT is one of EBM's most powerful tools. First and foremost, randomized controlled trials are experiments in which one group of patients receives the treatment being studied while a control group of patients receives either the standard treatment or no treatment at all (Macgill and Murrell 2018). The allocation of patients is done randomly in order to prevent the deliberate manipulation of results, although it will be ensured that common clinically relevant factors like age, sex, or ethnicity are evenly represented in both groups (Macgill and Murrell 2018). It is posited that by dividing groups in this way, other unknown but potentially influential variables will be equally distributed as well (Worrall 2010, 359). In short, the claim is made that since RCTs manage to balance all relevant variables between the two groups such that the only difference between them is whether or not they are receiving a particular treatment, it is the best way in which to measure the treatment effect.

While RCTs have significant methodological strengths, as outlined above, they are not without fault. One of the issues with RCTs is that they require researchers to use background knowledge in the randomization process to create groups that are balanced (Worrall 2010, 358). This means that researchers implicitly evoke hypotheses that may affect the outcome of the experiment. Thus, the main hypothesis (of whether or not a certain treatment is effective) cannot be tested in isolation. This problem is famously known as the Duhem-Quine thesis (Sankey 2019).

To illustrate this phenomenon, consider a researcher who chooses to randomize their population such that the control and treatment groups are balanced with respect to sex, age, and ethnicity. In this case, the researcher is implicitly hypothesizing that sex, age, and ethnicity are the main clinically relevant factors which may have an impact on how the treatment is received. However, it may be that after randomization one group happens to have more subjects who are smokers than the other. In this case, the study is not only assessing whether the treatment is effective, it is also implicitly testing the hypothesis that smoking is not clinically relevant to how a treatment, or lack thereof, affects a subject. This is just one (albeit extremely philosophical) way in which background knowledge informs and complicates the common understanding of RCTs.

More practically, researchers can often do quite well in balancing clinically relevant factors between two groups provided they have strong background knowledge of the condition they are investigating (Deaton and Cartwright 2018, 4). However, when background knowledge is incomplete, it is at least trivially possible that the two groups remain unbalanced in some significant way despite randomization (Worrall 2010, 358).

This insight was illustrated by a large RCT in which 3393 patients with bloodstream infections were randomized to receive either a prayer for their well-being and full recovery or no prayer at all (Leibovici 2001, 1450). The researchers en-

sured that the treatment and control groups were matched in all relevant factors. Having taken the standard precautions, it came as a great surprise that those in the prayer group's "length of stay in hospital and duration of fever were significantly shorter" (Leibovici 2001, 1450). Without a plausible mechanism in which prayer may act as a viable treatment for bloodstream infections, the researcher admitted that his results must have been influenced by a host of unknown factors and that the experiment was in fact a "non-study" (Worrall 2010, 359). In replying to his own work, Leibovici said that his goal was to encourage readers to question whether they would consider studies that were methodologically correct even if such a study lacked a plausible biological mechanism (Leibovici 2002). Although quite contrived, this high-profile case illustrates how evidence from RCTs cannot always be attributed solely to the treatment itself.

Once RCTs are published, it then becomes important for clinicians to translate that knowledge to their practice. The implicit belief is that trends observed in a sample group will carry over to patients provided that the studied group is sufficiently large and has similarities to the patients in question. Extrapolating a conclusion in this way is known as induction. However, EBM too often considers only a very small subsection of the population.

The National Institute of Health's Revitalization Act was passed in 1993 with the goal of increasing the number of women and racial minorities participating in research (Oh, 1). However, a 2015 study concluded that most healthcare practitioners and researchers are "informed by research extrapolated from a largely homogenous population, usually white and male" (Oh, 1). Less than one percent of cancer clinical trials focus on racial or ethnic minority populations, and minority populations were underrepresented in cancer research as a whole (Chen et al. 2014, 1093). Cardiovascular research suffers from a similar problem: less than a quarter of studies reported patient race while females made up just 30% of RCT samples (Sardar et al. 2014, 1868-69).

With practitioners of EBM likely treating a much more diverse group of patients, the argument of induction from clinical trials' evidence to general healthcare practice does not always hold. Research participant samples may not align with the characteristics of patients who require care, meaning that clinicians must be wary of transferring insights from research to their practice. In order to improve the applicability of findings, it is imperative that researchers work towards both reporting on and diversifying the make-up of their participant samples.

Here is reason to question EBM's faith in RCTs as the gold standard for evidence. This is not to say that other study designs are more or less appropriate in a clinical setting. However, if a healthcare practitioner adopts a treatment method because it has been proven effective by an RCT, and if they accept EBM's proposition that RCTs provide the best high-quality evidence, they may put more trust in their new protocol than is potentially warranted.

The Role of Statistical Significance in Evidence-Based Medicine

One of the ways that healthcare practitioners judge whether or not to follow a new treatment protocol is by looking to see if a study achieves statistical significance. In many medical journals, and in science more generally, statistical significance is commonly represented by a p-value that indicates the likelihood that a certain result could have occurred purely due to random chance. For example, a p-value of 0.025 would suggest that there was a 2.5% chance that the results of the study could have occurred randomly; in this case, the interpreter would conclude that the results were likely due to the effect of the treatment. But at what point is a p-value sufficiently small enough for researchers to decide that the treatment was effective or not? Generally, if the study returns a p-value of greater than 0.05, the result would not be considered statistically significant while a p-value of less

than 0.05 would indeed indicate a significant result.

This is problematic for a number of reasons. First, the threshold of $p \leq 0.05$ may incline healthcare practitioners to develop very different beliefs about a treatment based on whether or not it carries the label of statistical significance (Walsh et al. 2014, 623). In reality, a treatment with a p-value of 0.049 is probably not much more effective than a treatment with a p-value of 0.051. Additionally, it is possible that clinicians may put the same amount of faith in studies with similar p-values without considering other indicators of quality in a trial such as the number of events that occurred or the size of the control and intervention groups (Walsh et al. 2014, 623). Of course, not all clinicians may fall into this simplified way of thinking.

The broader issue here is that the p-value offers a dichotomous choice of significant vs. non-significant when it should instead present itself on a spectrum. The first time a threshold of $p \leq 0.05$ was suggested as a way to determine statistical significance was in Fisher's *Statistical Methods for Research Workers*, published in 1925 (Cowles and Davis 1982, 553). A year later, Fisher made this claim more explicit by saying that "personally, [he] prefers to set a low standard of significance at the 5 per cent point" although others may feel comfortable using a different standard (Fisher 1926, 504). Thus, the choice of which p-value threshold one uses is subjective. If pre-determined thresholds and the significant versus non-significant binary were removed, clinicians who evaluate the data would be better able to appraise the strength of a study's claims on their own terms.

Furthermore, recent research by Walsh et al. has showed that p-values in medical literature are not very stable. This phenomenon has been described by the Fragility Index, a measure that reports the minimum number of patients whose outcome would have to change in order for a statistically significant result to become non-significant (Walsh et al. 2014, 623). High scores are indicative of a robust study since the p-value would have risen above 0.05 only if a higher number of patients showed different outcomes. Lower scores suggest that the study's results are fragile, as the p-value would have risen above 0.05 if just a few patients had experienced different outcomes. In a systematic review examining 399 trials with a median sample size of 682 patients, it was found that the median Fragility Index was 8 (Walsh et al. 2014, 622). Remarkably, Walsh examined a study that enrolled 2316 patients and reported a p-value of 0.04, yet had a Fragility Index of 1; had just one of those 2316 patient experienced an unfavorable outcome, the study would have lost significance (Walsh et al. 2014, 626). In 53% of trials, the Fragility Index was less than the number of patients lost to follow-up, suggesting that if they had been retained, the treatment may not have been labelled as significantly effective (Walsh et al. 2014, 622).

Although more research needs to be done to properly judge whether medicine suffers from a widespread case of fragile results, these findings do indicate the need for healthcare professionals to look beyond the "statistically significant" label when deciding whether or not to alter their practice. It is thus very important for clinicians to look critically at studies before placing trust in them, if they do not do so already.

Having laid out some concerns with EBM from a philosophy of science perspective, two recommendations for improvement will now be provided.

1. Consider Epistemic Values

Given how RCTs and p-values can sometimes be misleading, healthcare practitioners and researchers could benefit from a discussion on epistemic values. The term "epistemic values" encompasses what is "acceptable in science as guidance for theory choice"; in this way, the values lay down the standards of evidence in a certain discipline (Douglas 2007, 120). In any clinical study, authors often decide how they

will interpret data in accordance with the model of study design they have chosen. However, researchers rarely discuss the reasons behind their methodological decisions to avoid colouring their paper with subjectivity (Douglas 2007, 123). This is not ideal since every choice that a researcher makes in the course of their study could potentially lead to error.

For example, imagine a scenario in which the marker of statistical significance is a p-value of less than 0.02. In this case, treatments would be held to a higher standard and the risk of discounting effective treatments would increase. In the current practice of EBM, the 0.05 threshold encourages practitioners to try potentially beneficial treatments even if they are more likely to fail (or do harm) than a 0.02 threshold would recommend. Clearly, the choice to adopt the 0.05 benchmark must have been the product of a value-based decision that prioritizes the potential benefits of novel action over the potential harms of inaction.

If researchers explicitly stated why they judged their evidence by certain standards and what epistemic values influenced their practical choice to designate a certain study as statistically significant, it would be easier for EBM practitioners to discuss how epistemic values influenced the confidence with which they applied a new treatment. This is especially important in cases where the treatment falls right on the edge of the threshold for statistical significance.

2. Strive for Epistemic Humility

Another way in which the field of EBM may be improved is by embracing the practice of expressing claims with epistemic humility. This means that healthcare professionals should strive to communicate to patients the degree to which the evidence suggests that a certain claim is accurate, or a treatment is effective (Schwab 2018, 29). This also helps to encourage healthcare practitioners to recognize the uncertainty in their field, and to nuance their professional judgement more carefully than the brutally simplistic dichotomous distinctions of statistical significance or RCT versus non-RCT. Additionally, it may be helpful for patients to become informed about p-values, the Fragility Index, and other common metrics of a study's strength so that they can make an informed choice regarding their treatment (Schwab 2018, 41). With epistemic humility being such a simple yet powerful addition to a healthcare practitioner's toolkit, it is undoubtedly a practice that should be more widely encouraged.

Conclusion

What started as a mutter among researchers at McMaster University has turned into a shout echoing across the world of healthcare: evidence-based medicine has come a long way and it is likely here to stay. Rather than decrying such a meteoric rise in prominence, this essay's philosophy of science analysis was meant to nuance the reader's understanding of current medical practice thus enabling them to navigate the field of EBM as a better-informed healthcare provider or patient. Traditional markers of quality evidence like RCTs and the statistically significant label were analysed, encouraging the reader to carefully examine available evidence in order to adopt the best treatment practices. It is my hope that with these things in mind, both patients and practitioners will be able to make carefully considered decisions and advocate for more openness in the research process.

Author Biography

Dinesh Moro is a student pursuing his undergraduate degree in Knowledge Integration at the University of Waterloo. His interests focus on understanding medicine as both a scientific and social process. Confronting the Obvious: An Epistemological Examination of the Evidence Informing Evidence-Based Medicinewas adapted from work produced as part of Dr. Kathryn Plaisance's

class on the Nature of Scientific Knowledge. It was inspired by the author's own research into the Fragility Index and a budding interest in the philosophy of science. The author is thankful for JIRR's cross-disciplinary focus and its commitment to elevating student writing.

Works Cited

Burns, Patricia B., Rod J. Rohrich, and Kevin C. Chung. "The levels of evidence and their role in evidence-based medicine." *Plastic and reconstructive surgery* vol. 128,1 (2011): 305-10.

Chalmers, A. F. *What Is This Thing Called Science?* Hackett Publishing Company, Inc., 2015.

Chen, Moon S. et al. "Twenty Years Post-NIH Revitalization Act: Enhancing Minority Participation In Clinical Trials (Empact): Laying The Groundwork For Improving Minority Clinical Trial Accrual." *Cancer* 120 (2014): 1091-1096. Web. 8 Apr. 2019.

Claridge, Jeffery A., and Timothy C. Fabian. "History and Development of Evidence-Based Medicine." World Journal of Surgery, vol. 29, no. 5, Springer Nature, Apr. 2005, pp. 547-553. Crossref, doi: 10.1007/s00268-005-7910-1.

Cohen, Aaron Michael, and William R. Hersh. "Criticisms Of Evidence–Based Medicine." *Evidence-based Cardiovascular Medicine* 8.3 (2004): 197-198. Web. 8 Apr. 2019.

Cowles, Michael, and Caroline Davis. "On The Origins Of The .05 Level Of Statistical Significance.." *American Psychologist* 37.5 (1982): 553-558. Web. 8 Apr. 2019.

Deaton, Angus, and Nancy Cartwright. "Understanding And Misunderstanding Randomized Controlled Trials." *Social Science & Medicine* 210 (2018): 2-21. Web. 8 Apr. 2019.

Douglas, Heather. *Rejecting The Ideal Of Value-Free Science.* Oxford University Press, 2007. Print.

Fisher, R. A. The arrangement of field experiments. *Journal of the Ministry of Agriculture*, 1926, 33, 503-513.

Godwin, Marshall, and Rachelle Seguin. "Critical Appraisal Skills Of Family Physicians In Ontario, Canada." *BMC Medical Education* 3.1 (2003): n. pag. Web. 9 Apr. 2019.

Goldenberg, Maya J. "On Evidence and Evidence-Based Medicine: Lessons from the Philosophy of Science." Social Science & Medicine, vol. 62, no. 11, Elsevier BV, June 2006, pp. 2621-2632. Crossref, doi:10.1016/j.socscimed.2005.11.031

Greenhalgh, T., J. Howick, and N. Maskrey. "Evidence Based Medicine: A Movement In Crisis?." *BMJ* 348.jun13 4 (2014): 1-7. Web. 8 Apr. 2019.

Haynes, R Brian. "What Kind Of Evidence Is It That Evidence-Based Medicine Advocates Want Health Care Providers And Consumers To Pay Attention To?." *BMC Health Services Research* 2.1 (2002): n. pag. Web. 8 Apr. 2019.

Leibovici, Leonard. "Effects of Remote, Retroactive Intercessory Prayer on Outcomes in Patients with Bloodstream Infection: Randomised Controlled Trial." BMJ, vol. 323, no.7372, BMJ, Dec. 2001, pp.1450-1451. Crossref, doi:10.1136/bmj.323.7372.1450.

Leibovici, Leonard. "Authors' Reply." *(*2002): n. pag. Print.

MacGill, Markus, and Daniel Murrell. "Randomized Controlled Trials: Overview, Benefits, And Limitations." *Medical News Today.* N.p., 2018. Web. 11 Dec. 2018.

Oh, Sam S et al. "Diversity in Clinical and Biomedical Research: A Promise Yet to Be Fulfilled." *PLoS medicine* vol. 12,12 e1001918. 15 Dec. 2015, doi:10.1371/journal.pmed.1001918

Plaisance, Kathryn. "Holism" INTEG 220. 04 Oct. 2018, Waterloo, University of Waterloo.

Sackett David L, Rosenberg William M C, Gray J A Muir, Haynes R Brian, Richardson W Scott. "Evidence based medicine: what it is and what it isn't" *BMJ* 1996; 312 :71

Sankey, Howard. "Quine-Duhem Thesis." *Philpapers.org.* N.p., 2019. Web. 8 Apr. 2019.

Sardar, Muhammad Rizwan et al. "Underrepresentation Of Women, Elderly Patients, And Racial Minorities In The Randomized Trials Used For Cardiovascular Guidelines." *JAMA Internal Medicine* 174.11 (2014): 1868. Web. 8 Apr. 2019.

Schünemann, Holger J., et al. "Improving the Use of Research Evidence in Guideline Development: 9. Grading Evidence and Recommendations." Health Research Policy and Systems, vol. 4, no. 1, Springer Nature, Dec. 2006. Crossref, doi:10.1186/1478-4505-4-21.

Schwab, A. "Epistemic Humility And Medical Practice: Translating Epistemic Categories Into Ethical Obligations." *Journal of Medicine and Philosophy* 37.1 (2012): 28-48. Web. 12 Dec. 2018.

Timmermans, Stefan, and Aaron Mauck. "The Promises And Pitfalls Of Evidence-Based Medicine." *Health Affairs* 24.1 (2005): 18-28. Web. 8 Apr. 2019.

Tinker, Ann. "The Top 7 Healthcare Outcomes Measures." *Health Catalyst.* N.p., 2018. Web. 9 Dec. 2018.

Walsh, Michael, et al. "The Statistical Significance of Randomized Controlled Trial Results is Frequently Fragile: A Case for a Fragility Index." Journal of Clinical Epidemiology, vol. 67, no. 6, Elsevier BV, June 2014, pp. 622-628. Crossref, doi:10.1016/j.jclinepi.2013.10.019

Williams, Ian A. "Cultural Differences In Academic Discourse: Evidence From First-Person Verb Use In The Methods Sections Of Medical Research Articles." *International Journal of Corpus Linguistics* 15.2 (2010): 214-239. Web. 8 Apr. 2019.

Wilson, H. J. "The Myth Of Objectivity: Is Medicine Moving Towards A Social Constructivist Medical Paradigm?." *Family Practice* 17.2 (2000): 203-209. Web. 8 Apr. 2019.

Worrall, John. "Evidence: Philosophy of Science Meets Medicine." Journal of Evolution in Clinical Practice, vol. 16, no. 2, Wiley, Mar. 2010, pp. 356-362. Crossref, doi:10.1111/j.1365-2753.2010.01400.x.