The judgements that evidence-based medicine adopts

Elena Rocca

Abstract

In 'The evidence that evidence-based medicine omits', Brendan Clarke and colleagues argue that when establishing causal facts in medicine, evidence of mechanisms ought to be included alongside evidence of correlation. One of the reasons they provide is that correlations can be spurious and generated by unknown confounding variables. A causal mechanism can provide a plausible explanation for the correlation, and the absence of such an explanation is an indication that the correlation is not causal. Evidence-based medicine (EBM) proponents remain sceptical about this argument, one problem being that the formulation of a mechanism requires judgements that are external to the evaluation of data and experimental designs - for instance judgements of plausibility against, or derivability from, background knowledge. Since background knowledge is always incomplete and therefore unreliable, EBM proponents maintain that the plausibility of a hypothesis should be evaluated mainly by the quality of population data that yielded it. Here, I use the example of oestrogen replacement therapy's effect on coronary heart disease, an example that is often quoted in defence of the epistemic advantage of randomised controlled trials, to show that the evaluation of the most reliable study design necessarily implies the adoption of judgements that are external to the specific evidence of correlation. The exclusion of evidence of mechanism, therefore, is not effective in bypassing paradigm-dependent judgements, which are external to specific evidence. Since such judgements cannot be excluded by evidence evaluation, they can only be kept under scrutiny, or adopted uncritically. I propose that the latter option can hinder the maintenance of an active critical inquiry, as well as the analysis of experts' disagreement

Keywords: Evidence evaluation, statistical evidence, mechanism, medicine, decision-making.

1. Introduction

The approach to clinical practice proposed by the evidence-based medicine (EBM) movement, has been a topic of debate during the last few decades. EBM is defined by its proponents as 'the integration of best research evidence with clinical expertise and patient values' to guide the clinical care of the single patient (1). While highlighting that evidence alone is insufficient to guide clinical decisions, this definition assigns nevertheless a crucial role to 'the best research evidence'. However, there is debate on the *kind* of research evidence that best supports decision-making in clinical practice. It is generally accepted that the establishment of statistical correlations through comparative population studies can be a powerful tool for detecting causal relationships in medicine. Older age (over 50) is correlated with an increased onset of colorectal cancer, for instance, which suggests a causal role of aging in the aetiology of this condition, and grounds the promotion of screening programs for this age group by some health care systems. However, it is also generally accepted that correlations are only indicators of causation, and thus can be fallible. In the previous example, age might not be a genuine cause of colorectal cancer, but just be associated with it because the two might be caused by common molecular changes. Even more problematically, correlations can be spurious, and be generated by unknown confounding variables.

In presence of evidence of correlation, how should one evaluate whether it is a genuine or a spurious one? One strategy that has been supported by some scholars in philosophy of medicine is to look for evidence of a plausible mechanism underlying the correlation (2) (3). While there are several definitions of 'mechanism' available in the philosophical and medical literature, it is beyond the scope of this article to commit to any one particular definition. For my purposes here, it will be sufficient to generally define a mechanism as a causal process underlying the correlation. Evidence addressing a causal mechanism can include qualitative observations, laboratory experiments, investigations at the micro level, as well as existing statistical studies (4). The role of a causal mechanism is to provide a plausible explanation for the correlation, and the absence of such an explanation is an indication that the correlation is spurious (2).

In tension with this argument, EBM frameworks rate evidence of correlations higher than the evidence of underlying mechanisms (5). That is to say, in the presence of high quality population studies, and therefore of reliable correlations, evidence of the mechanism underlying such correlations is not deemed as necessary for supporting decision making (6). Instead of collecting different types of evidence, EBM proponents aim to improve the reliability of a correlation by improving the quality of the population studies that yield it:

> '…quality (tightly controlled, unbiased etc.) evidence rather than quantity of evidence helps reduce the likelihood of spuriousness.' (6).

Why are EBM proponents sceptical about the reliance on causal mechanisms for establishing a causal relationship? One problem they wish to bypass can be illustrated as follows. Imagine that, for the formulation of a certain mechanism M, one needs three sets of specific data, for example, post-mortem histology, genotyping data, and cell-culture experiments. In order for these data to count as evidence for the specific mechanism M, they need to corroborate a number of related research hypotheses. In addition, the experimental design and methods used to collect and analyse the data ought to be both relevant and reliable. In the following, I will call *specific evidence* the set of data, together with the methods for data collection and analysis, which are used to directly test a *specific* hypothesis. Any other information will be called background knowledge. This includes more general evidence as well as theoretical understanding.

However, the evaluation of data and experimental designs are not sufficient for the purpose of establishing M. First, the mechanism will not be established if it is not judged as 'plausible' based on background knowledge, which is external to the specific evidence. For instance, when investigating the mechanism underlying allergic reactions to a certain drug, the specific evidence will normally be interpreted in accordance with the basic knowledge about the physiology of the immune system. Such established knowledge is itself based on evidence, however this is a previously accumulated, general evidence, rather than a specific one, directly addressing the research question. This reliance on plausibility is problematic since background knowledge is always incomplete, and any judgement that leans on it is therefore unreliable. In multiple historical instances, based on commonly accepted background knowledge, the scientific community rejected the correct mechanism as implausible, or accepted the wrong mechanism as plausible (4).

> 'The apparent knowledge of what happens to some of the mechanisms under intervention lends an aura of acceptability, which, in turn, leads to more prolific use of a harmful effect' (6, p.934)

Second, the mechanism will not be established if it is not derived from within a more general understanding of phenomena, which is also external to the specific evidence. The formulation of the mechanism underlying a certain drug intervention, for example, requires an understanding of general mechanisms of drug metabolism. This is problematic for the same reason as above: such understanding is always incomplete:

> 'When it comes to drug therapy, mechanistic reasoning is bound to be based on 'partial' mechanisms because of the complexity and somewhat mysterious metabolic mechanism' (7).

We see, therefore, that the establishment of a causal mechanism, or explanation, requires additional judgements respects to the mere evaluation of the specific evidence (data and

experimental designs). These judgements are never entirely reliable, because they are based on background knowledge and general understanding of phenomena, which are always incomplete.

In response to this, EBM frameworks recommend to rely preponderantly on evidence of correlations from population data, when making a causal claim (6). This recommendation can be motivated as follows (6, 7). While evidence of mechanism needs to be interpreted against fallible background knowledge and theoretical understanding, population data can be graded by simply evaluating the experimental design that generated them. This process still requires some theory (statistical theory, for instance). However, it largely circumvents the reliance on judgements that are external to the evidence of correlation available to test the specific hypothesis. Thus, evaluating the quality of a population study is comparatively a straightforward matter, since our understanding of what constitutes (for instance) a good statistical design is more reliable than our general understanding of biological phenomena. For this reason, according to the EBM framework, good evidence of correlation from population studies should be normally given priority over mechanistic evidence in medical decision making.

In this paper, I argue that the strategy proposed by EBM, of relying preponderantly on correlations from population data, is not successful for the purpose of bypassing judgements that are external to the specific evidence. When the explanation underlying a certain correlation is left aside, the relative quality of population studies is proposed as the only tool in order to evaluate a casual hypothesis (6). This process, however, cannot be done by considering exclusively the population studies themselves; it is still reliant on external judgements based on background understanding of phenomena. The normative divergence, therefore, is not whether judgements external to the specific evidence ought to be included in medical decision making, but whether they ought to be kept under scrutiny or accepted uncritically (fig. 1).

2. Weighing evidence from different population studies

At times, population studies yield correlations that contradict previous evidence or established theories. They might even contradict intuitions that are more fundamental than established theories. A randomised controlled trial (RCT), for instance, tested the effect of retroactive prayer on the length of hospitalization of 3400 patients and found a statistical significant effect of the 'treatment' in the experimental group (8). The author's motivation for this experiment was to show that a good experimental design is not a guarantee for unbiased results. In other words, the author adopts the fundamental assumption that instances of backward causation can never exist, therefore any result pointing to the opposite conclusion must be wrong. On the other hand, if one disregards the primitive role of basic assumptions, the results of such a study would give reasons to doubt the basic knowledge that the cause always precedes the effect. Are correlations of this kind spurious, or do they indicate real causation? EBM proponents advise

caution in evaluating the plausibility of these hypotheses by looking for an explanatory causal mechanism based on current background knowledge. In other words, we should not rely on the fact that basic knowledge does not provide us with a plausible theory that would explain backward causation. We would risk repeating historical errors, in which relevant evidence was dismissed because at the time there was no established causal mechanism that could make sense of it. As an example, in the much quoted case of Dr. Semmelweis, evidence that antiseptic routines reduce infections at childbirth was rejected as implausible because there was no accepted understanding of how this could happen (9).

How can such mistakes be avoided? Is there a strategy that allows to evaluate the spuriousness of a correlation, without relying on fallible mechanistic thinking? There is, according to EBM proponents. One can judge the plausibility of the hypothesis by evaluating the quality of the studies that yielded it. Howick 2011 states that it is possible to

> '[…] evaluate the comparative clinical studies *on their own grounds*[1]. Implausible hypothesis are either true or false. If true we would expect consistent detectable effects in unbiased comparative clinical studies' (6, p. 932, emphasis mine).

It is the presence or absence of such effects, without the need of any explanation (and therefore without the need of any additional judgement than the evaluation of the study itself), that should guide our evaluation of the hypothesis. Ideally then, postulates Howick, if unbiased comparable clinical studies showed consistently that retrospective prayer improves the healing of hospitalized patients, we ought to start believing that the cause can come after the effect, at least sometimes(6).

> 'If such consistent effects are demonstrated, then we should recall the Semmelweis case and temper our skepticism regarding the plausibility of the hypothesis' (6, p. 932)

In other words, the quality of the evidence alone is the most reliable benchmark to evaluate the likelihood of a hypothesis.

This type of argumentation relies on the assumption that the evaluation of the relative strength of evidence, the quality of evidence and bias do not adopt judgements that are external to the specific evidence of correlation. Only under this assumption, indeed, can such evaluations be considered independent from background knowledge, explanations, and broader understanding of biological phenomena and therefore more reliable that mechanistic thinking.

This assumption, I will argue, might hold in some simple cases of evidence evaluation, for instance when a trial with considerable statistical power stands against a previous small study. However,

[1] By 'on their own grounds' the author means without the need of mechanistic reasoning. He defines mechanistic reasoning as 'an inferential chain (or web) linking the intervention […] with a patient-relevant outcome, via relevant mechanism'. In other words, mechanistic reasoning is establishing causation by finding an explanation for why and how the outcome happens.

it cannot be generalized to evidence evaluation overall and therefore should not be taken as the default assumption. To demonstrate my claim, I will consider a renowned example within the EBM discourse: the case of oestrogen replacement therapy and its effect on coronary heart disease.

Up until the 1990s, oestrogen therapy was widely prescribed to post-menopausal women because it was believed to protect them from coronary heart disease. Besides the existence of mechanistic evidence, this belief was supported by a considerable body of observational population studies. Although not all the observational evidence was consistent, much of it pointed to the same results. For instance, 15 internally controlled, medium and large prospective observational studies showed consistently that post-menopausal women receiving oestrogen replacement therapy had significantly less incidence of coronary heart disease than untreated patients (overall relative risk 0.50, confidence interval 0.43-0.56) (10). The confidence in oestrogen replacement therapy, however, was suspended when a large RCT showed the opposite result: treated patients had significantly increased risk of heart disease (estimated hazard ratio 1.29, confidence interval 1.02-1.63) (11).This example is often quoted to demonstrate the epistemic advantage of RCTs over other types of evidence. However, it is circular to affirm that the RCT was considered superior because it yielded the 'true' result (6). In a weight-of-evidence approach, which considerations do we need to adopt in order to evaluate the RCT as the most reliable evidence? Let us consider three steps that are needed in order to evaluate the quality of the RCT against the existing body of evidence.

A) *Quality of the randomisation (known confounders)*. The EBM framework adopts the a priori assumption that causation is best detected by its capacity to make a difference. One method to test difference-making is to apply a potential causal factor to only one of two equivalent settings, so that the difference between the two outcomes can be attributed to the tested factor. The essential prerequisite is that the two settings - the experimental setting to which the tested causal factor is applied and the control setting that remains untreated - must be as similar as possible in order to circumvent problems such as causal over-determination and/or masking of the effect by unbalanced causal relevant factors (confounders). In order to meet this requirement, in a clinical trial the group of patients receiving the treatment must be as 'similar' as possible to the control group. This corresponds to saying that patient's characteristics that are causally relevant for the outcome must be equally distributed among the two groups.

In observational studies, the task of equally distributing the potentially relevant causal factors among the two groups is done by 'matching' the experimental and control group for the factors that one thinks might be significant. The clear problem with this strategy is that there is an indefinite number of unknown relevant factors, over which we have no control. This problem is supposedly taken care of by an alternative strategy, randomisation. The reasoning here is that if

one randomly assigns the patients to the two test groups a priori and the groups are large enough, then there is greater probability that more known and unknown confounders will be distributed evenly across the group, compared to observational studies (12). However, random allocation together with the law of large numbers is not *per se* sufficient reason to be confident that the study groups are more balanced than observational studies of otherwise comparable quality. At least not according to EBM textbooks. Sackett and colleagues do not take randomisation to be a guarantee for a high level of homogeneity without a so-called 'baseline assessment'. This consists in a 'double-check to see whether randomisation was effective by looking to see whether patients were similar at the start of the trial' (12). If randomisation did not work properly, then it is necessary to re-randomise the participants.

Verifying whether 'patients are similar' means double-checking whether experimental and control groups have a similar distribution of confounders that we suspect will be causally relevant, or factors that 'potentially affect the outcome' (7). What is to be accounted as potentially relevant factors change depending on the case at stake and, crucially, cannot be derived by the specific data of correlation themselves. On the contrary, it is a judgement based on a broader background understanding of the phenomenon under scrutiny. In the Rossouw RCT, the baseline assessment included age, race, hormone use, body mass index, smoking, history of heart disease, treatment for diabetes, blood pressure, statins and aspirin use amongst other factors. As it should be clear, such selection of causally relevant factors is reliant on the available body of knowledge about heart disease aetiology. This body of knowledge (obviously) varies over time. In the 1970s, for example, it was already clear that heart disease was correlated with social class distribution, but by 1992 a partial explanation was offered for this correlation: different levels of education contributed to the level of cigarette smoking, blood pressure, and total cholesterol (13) (14). The evaluation of potentially relevant causal factors, therefore, would have been different at these two different time points. When evaluating whether randomisation was effective, hence the quality of the RCT, hence the most reliable result, one unavoidably adopts the contemporary advancement in understanding the related phenomena.

B) *Quality of randomisation (unknown confounders)*.

Although RCTs maintain a primary epistemic role in EBM, the original claim that a well-done RCT trumps observational evidence, no matter how good and how much of it (15), was harshly criticized (16). One problem is that although we verified that randomisation was effective for known confounders, we still cannot be sure about how it took care of all the unknown ones. As generally acknowledged (by EBM proponents as well), randomisation alone is not sufficient to assure that all confounding factors are equally distributed across groups, even when it guarantees high probability of balancing each single factor (7,8,17,18). Some scholars have even proposed that 'there is good reason to doubt that the balance assumption [the assumption that

each potential confounding cause is distributed similarly among the study groups] is true for even one of our best RCTs' (18). Is this a matter of concern? It could be if we endorse arguments defending that RCTs should control for all causally relevant factors (19). However, in the case at hand, we do not necessarily need to ensure that the Russow study controlled for all confounding factors, as long as we demonstrate that it controlled for more confounding factors than the previously available observational evidence.

This, again, is hard to do by solely focusing on the study designs and results. By adhering to these elements, one can only observe that the Rossouw study, because it is randomised, is the only study that removes allocation biases, therefore it is, in principle, the study that has better chances of being the least biased. This speculation can be proven much more firmly, however, by adopting judgements external to the studies themselves. The same Howick 2011 provides us with an example of such judgements:

> '…there are independent reasons to believe the results of the randomised trial […]. For example, the authors of the earlier studies observed that mortality due to homicide was higher among women who did not take HRT [oestrogen replacement]. This (among other similar observations) implied that there were potentially confounding differences between women who chose (or were chosen by their doctors) to take HRT and those who did not.' (6, p. 932)

What Howick suggests here, is to evaluate the observational studies by considering some of the correlations they yield and judge them by their plausibility against background knowledge, rather than by their statistical significance. We do not accept that oestrogen replacement protects women from death due to homicide. Rather, we accept that this correlation points out an imbalance in the study recruitment, whereby one of the groups had a higher number of women living in bad neighbourhoods, for instance. These considerations, however, would not be possible if we focused only on the study in isolation. In this case, indeed, we would only be allowed to consider, as benchmarks for the plausibility of the causal association, factors such as the degree of significance, and/or the magnitude of the effect. The bigger the latter two, the more plausible it would be that the correlation is not spurious. When external judgement (based on general understanding of biology) is instead adopted, the stronger one such correlation, the bigger the proof that the study is seriously biased.

C) *Analogy with existing population studies.*

One could argue that, although considerations external to specific evidence might be necessary for evaluating the relative strength of population studies, such considerations do not need to include theoretical – explanatory aspects. Specifically, the reliance on a study's result can increase if the study is in line with other high quality population studies. For example, by the early 1970s there was evidence that the administration of equine oestrogen did not protect men

against heart coronary disease. On the contrary, high doses of oestrogen were correlated with a higher incidence of the condition (20). These results, yielded by the large cohort 'Coronary Drug Project', were used as an argument for holding scepticism toward the positive outcomes of observational trials in women (21). This way, one could argue that every new population study might be evaluated in the context of previously existing high-level population studies. In other words, background knowledge might be limited to population data, and exclude fallible explanations.

The trouble with this reasoning is that, although in principle it is possible to separate population studies from the general state of the pathophysiological understanding of phenomena, this is not a realistic account of how evidence evaluation actually happens. Evidence from population studies, like any other type of evidence, acquire different significance depending on the *total* body of knowledge already available. For instance, gender difference in coronary heart disease have been increasingly elucidated, disclosing different anatomy, different physiopathology, different response to therapy and gender-specific risk factors (22). While in the 1990s the gender – specific outcome of a cardiovascular intervention might have been surprising, it is less surprising nowadays. Arguably therefore, the fact that oestrogen replacement did not protect men from heart conditions would be less persuasive as an argument for scepticism toward the positive effect in women. By drawing on background knowledge, judgements external to the specific evidence also draw on explanatory, theoretical understanding of phenomena.

In summary, I have considered the case of oestrogen replacement therapy and coronary heart disease, in which population studies of comparable statistical power, but with different experimental design, yielded conflicting results. When referring to this and similar examples, EBM proponents urge that the quest of an explanatory mechanism by which oestrogen might protect from heart disease, or on the contrary provoke it, should not have a decisive role in the evaluation or total evidence. On the contrary, they suggest, the evaluation of the most reliable result ought to exclude judgements that are external to the population studies themselves. In this section, I have argued that the evaluation of which experimental design was to be considered the most reliable, and therefore which result ought to be taken as 'true', inherently implies the use of judgements that are external to the specific population studies. Indeed, in order to justify the epistemic advantage of the randomised study, one needs to be confident in the fact that randomisation worked. This, as I suggested, cannot be done with the use of specific evidence alone, but drawing on background knowledge and the understanding of phenomena. The strategy of limiting evidence to correlations from population studies, therefore, is not effective in bypassing the need of judgements that are external to specific evidence.

3. Implications

If judgements that are external to the specific evidence are inherently built into the evaluation of conflicting results, one cannot decide whether or not to include them in medical decision-making. The choice is rather between keeping them under scrutiny, or accepting them uncritically.

EBM's high reliance on population studies 'on their own', and the low epistemic status left for explanatory mechanisms, conveys the false impression that specific evidence alone can guide us toward the 'correct' evaluation, and ultimately toward the right decision. There are some risks, or at least some disadvantages, in promoting such a belief on the supremacy of data. For one, it is not uncommon that good data lead to the wrong answer. For instance, much evidence in the 1970s led the scientific community to wrongly accept herpes virus (HSV) as the cause of cervical cancer (4). In a study on 40.000 screened patients, HSV was strongly correlated with malignant changes of the cervical epithelium (23). Cervical dysplasia was correlated with HSV2 infections (24). Antibodies against HSV2 were present four times as often in women with cervical cancer as controls (25). Fragments of HSV DNA could be directly detected in cervical cancer cells (26). Other types of herpes viruses were also implicated in other types of cancer (4). The wrong causal connection between HSV and the aetiology of cervical cancer was dominant until the mid-1980s. Only then, did the role of papilloma virus start to become evident.

This and a plethora of similar stories highlight that in order to maintain active the critical inquiry, any evidence needs to be met with scepticism, rather than belief (27). This becomes somewhat harder, under the illusion of a completely objective, data-driven analysis. On the contrary, if all types of judgements used in evidence evaluation are kept under scrutiny, it becomes easier to suspend belief.

Another problem with the illusion of a paradigm-free evidence evaluation is that it leaves us with no tools to understand cases of expert disagreement, in which different evaluations are drawn from the same evidence (28). There is rarely a straight forward path to decision making that is paved by the perfect data. Studies often yield conflicting results and have strengths and weaknesses that need to be weighed against each other - a process that necessitates expert's judgement.

Consider one example that has been increasingly visible in the last few years. The correlation between exposure to the herbicide glyphosate and the onset of some types of cancer, including non-Hodgkin lymphoma, was tested with a number of population studies. Three of these investigations are unanimously considered 'key studies' by different evaluating agencies because of large numbers and high statistical power (29). However, all these studies are also unanimously acknowledged to have weaknesses: low number of exposed cases, potentially unadjusted exposure to other pesticides, and possible biases in assessing the exposure (29). Two of these three studies are retrospective pooled analyses of case-control studies, one performed across Canada, and the other in the mid-west US (30,31). The specific advantage of these two studies is

that they control for confounding by other pesticides using a sophisticated statistical analysis. Both studies showed a positive correlation between exposure to glyphosate and the onset of non-Hodgkin lymphoma, while indeed controlling for several confounding exposures (32). However, it is generally acknowledged that such study design is prone to recall biases, since it is retrospective. On the other hand, a relatively large cohort study, the Agricultural Health Study (AHS), showed no statistical significant association between exposure to glyphosate and the onset of non-Hodgkin lymphoma, with no apparent exposure–response relationship in the results (33). There are potential advantages of cohort versus case-control studies: the cohort study is a prospective study and is hence free from recall biases. Despite this potential advantage, in this case the follow-up period was limited to less than a decade, which is considered an insufficient time frame for the onset of the lymphoma, at least according to the current understanding of this type of cancer (29).

As often happens, different agencies evaluated this epidemiological evidence in different ways. The European Food Safety Authority (EFSA), for instance, considered the cohort AHS study as the best epidemiological evidence available, as it is the only one that was able control recall biases. Accordingly, EFSA's verdict is that epidemiological data give 'very limited evidence' for an association between glyphosate – based formulations and non-Hodgkin lymphoma, 'overall inconclusive for a causal or clear associative relationship' (34). The International Agency for Research on Cancer (IARC), on the contrary, evaluated the two case-control studies as more reliable than the cohort study, one of the motivations being that 'the median follow-up time in the AHS was 6.7 years, which is unlikely to be long enough to account for cancer latency' (29). Consequently, IARC concludes that 'a positive association *has* been observed', although 'chance, bias or confounding could not be ruled out with reasonable confidence' (32, emphasis mine). This difference in evaluation is not as minimal as it seems. IARC, indeed, urges that 'legitimate public health concerns arise when causality is credible' (29), while EFSA finds no reason for such public concern (35).

If we focus solely on the population studies and the experimental designs, we are left with little ground to understand the foundations of this disagreement. Indeed, all agencies agree about what the strength and weaknesses of each experiment are. What changes from agency to agency is the evaluation of the relevance of such strength and weaknesses in relation to the current knowledge advance on cancer aetiology (for instance, when considering the length of follow up). Moreover, there is arguably a different value judgement of whether a potential false positive due to re-collection biases is preferable to a possible false negative due to insufficient study duration. The glyphosate case is at the moment an open controversy, showing once more that the mere evaluation of studies 'on their own grounds' poorly applies to the complex realm of health sciences. Overall, the proposition of a weight-of-evidence process, driven by population studies and excluding external judgements, seems more an ideal abstraction than a concrete solution.

In summary, I argued in favour of a pluralistic approach to causal evidencing in medicine. My argument in line with other stances advocating that both evidence of correlations and explanatory evidence are necessary for successful decision-making (2-4). I allow that, in some cases, *good* evidence from population studies can license a causal inference. The crucial point, however, is how one should assess that such evidence is 'good', or 'better' than evidence from other population studies, which might point to different conclusions. I showed a case in which such judgement presupposes theories of mechanisms. Arguably, this is representative for many other complex cases of causal inference. However, even assuming that my conclusion cannot be extended to the majority of evaluations of statistical evidence, it still shows how a rigid interpretation of EBM evidence hierarchy is problematic. Finally, I fully acknowledge that explanations can be wrong, and consequently can hinder the correct causal inference. However, since, as I argued here, such explanations are irreducibly embedded in the medical sciences, I see this fallibility as a motivation for increasing our enquiry on causal explanations, rather than for dismissing it.

References

1. Sackett DL, Straus SE, Richardson WS, et al. *Evidence-Based Medicine: How To Practice And Teach EBM*. 2nd ed. Edinburgh: Churchill Livingstone; 2000.
2. Russo F, Williamson J. Interpreting causality in the health sciences. *Int Stud Philos Sci.* 2007; 21:157–70.
3. Rocca E, Anjum RL, Mumford S. causal insights for failure. Post-marketing risk assessment of drugs as a way to uncover causal mechanisms. In: La Caze, A, Osimani, B, eds *Uncertainty in Pharmacology: Epistemology, Methods and Decisions*. Springer. Boston Series for the History and Philosophy of Science; 2018. Forthcoming.
4. Clarke BO. *Causality in Medicine With Particular Reference to The Viral Causation of Cancers*. 2011.
5. OCEBM Levels of Evidence Working Group. The Oxford 2011 Levels of Evidence. 2011.
6. Howick J. Exposing the vanities - and a qualified defense - of mechanistic reasoning in health care decision making. *Philos Sci*. 2011;78:926–40.
7. Howick J. *The Philosophy of Evidence-Based Medicine*. Oxford: Wiley-Blackwell, BMJ Books; 2011.
8. Leibovici L. Effects of remote, retroactive intercessory prayer on outcomes in patients with bloodstream infection: randomised controlled trial. *BMJ*. 2001; 323:1450–1.

9. Scholl R. Causal inference, mechanisms, and the Semmelweis case. *Stud Hist Philos Sci Part A*. 2013; 44:66–76.

10. Stampfer MJ, Colditz GA. Estrogen replacement therapy and coronary heart disease: A quantitative assessment of the epidemiologic evidence. *Int J Epidemiol*. 2004; 33:445–53.

11. Rossouw JE, Anderson GL, Prentice RL, et al. Risks and benefits of estrogen plus progestin in healthy postmenopausal women. *J Am Med Assoc*. 2002; 288:321–33.

12. Sackett D, Richardson W, Rosenberg W, Haynes B. *Evidence-Based Medicine: How To Practice And Teach EBM*. London: Churchill; 1997.

13. Marmot MG, Adelstein AM, Robinson N, Rose GA. Changing social-class distribution of heart disease. *BMJ*. 1978; 2: 1109–12.

14. Winkleby MA, Jatulis DE, Frank E, Fortmann SP. Socioeconomic status and health: How education, income, and occupation contribute to risk factors for cardiovascular disease. *Am J Public Health*. 1992; 82:816–20.

15. Barton S. Which clinical studies provide the best evidence? The best RCT still trumps the best observational study. *BMJ*. 2000; 321:255–6.

16. Rawlins MD. *De Testimonio: On The Evidence For Decisions About The Use Of Therapeutic Interventions*. London: The Harveian Oration; 2008.

17. Worrall J. Evidence: philosophy of science meets medicine. *J Eval Clin Pract*. 2010; 16:356–62.

18. Fuller J. The confounding question of confounding causes in randomized trials. *Br J Philosphy Sci*. 2018; 0:1–26.

19. Cartwright N. *Nature's Capacities and Their Measurements*. Oxford: Oxford University Press; 1989.

20. Group TCDPR. The Coronary Drug Project. Findings leading to discontinuation of the 2.5-mg day estrogen group. *JAMA*. 1973; 226:652–7.

21. Petitti D. Commentary: Hormone replacement therapy and coronary heart disease: Four lessons. *Int J Epidemiol*. 2004; 33:461–3.

22. Yahagi K, Davis HR, Arbustini E, Virmani R. Sex differences in coronary artery disease: Pathological observations. *Atherosclerosis*. 2015; 239:260–7.

23. Naib ZM, Nahmias AJ, Josey WE. Cytology and hystopathology of cervical herpex simplex infection. *Cancer*. 1966; 19:1026–31.

24. Cleator G, Klapper P. Herpes Simplex. In: Zuckerman AJ, Banatvala JE, Pattison JR, Griffiths P, Schoub B, eds. *Principles and Practice of Clinical Virology*. 5th ed. Chinchester: John Wiley & Sons; 2004: 7–51.

25. Rawls WE, Tompkins WAF, Figueroa ME, Melnick JL. Herpesvirus Type 2: Association with Carcinoma of the Cervix. *Science*. 1968; 161:1255–6.

26. Frenkel N, Roizman B, Cassai E, Nahmias A. A DNA fragment of herpex simplex and its transcription in human cervical cancer tissue. *Proc Natl Acad Sci U S A*. 1972; 69:3784–9.

27. Weed DL. Causal criteria of Popperian refutation. In: Rothman K, ed. *Causal Inference*. Chestnut Hill MA: Epidemiology Resources Inc.; 1988: 26.

28. Rocca E, Andersen F. How biological background assumptions influence scientific risk evaluation of stacked genetically modified plants : an analysis of research hypotheses and argumentations. *Life Sci Soc Policy*. 2017; DOI 10.118.

29. Portier CJ, Armstrong BK, Baguley BC, et al. Differences in the carcinogenic evaluation of glyphosate between the International Agency for Research on Cancer (IARC) and the European Food Safety Authority (EFSA). *J Epidemiol Community Heal*. 2016; 0(0):1–5.

30. McDuffie HH, Pahwa P, McLaughlin JR, et al. Non-Hodgkin's lymphoma and specific pesticides exposures in men: cross-Canada study of pesticides and health. *Cancer Epidemiol Biomarkers Prev*. 2001; 10:1155–63.

31. De Roos AJ, Zahm SH, Cantor KP, et al. Integrative assessment of multiple pesticides as risk factors fornon-Hodgkin's lymphoma among men. *Occup Environ Med* . 2003; 60:E11. Available from: http://oem.bmj.com/cgi/doi/10.1136/oem.60.9.e11

32. IARC Working Group. Glyphosate. In: *Some Organophosphate Insecticides and Herbicides: Diazinon, Glyphosate, Malathion, Parathion, And Tetrachlorvinphos Vol 112*. Lyon: International Agency for Research on Cancer; 2015: 321-99.

33. De Roos AJ, Blair A, Rusiecki JA, et al. Cancer incidence among glyphosate-exposed pesticide applicators in the Agricultural Health Study. *Environ Health Perspect*. 2005; 113:49–54.

34. European Food Safety Authority. Conclusion on the peer review of the pesticide risk assessment of the active substance glyphosate. *EFSA J.* 2015; 13:4302.

35. European Food Safety Authority. *Final Addendum to the Renewal Assessment Report*. 2015.