Corroborating Evidence-Based Medicine

(Please cite the published version, forthcoming in *Journal of Evaluation in Clinical Practice)*

Alexander Mebius

Department of Philosophy and History of Technology, Royal Institute of Technology (KTH), Teknikringen 78 B, 100 44 Stockholm, Sweden

Abstract

Proponents of evidence-based medicine (EBM) have argued convincingly for applying this scientific method to medicine. However, the current methodological framework of the EBM movement has recently been called into question, especially in epidemiology and the philosophy of science. The debate has focused on whether the methodology of randomized controlled trials provides the best evidence available. This paper attempts to shift the focus of the debate by arguing that clinical reasoning involves a patchwork of evidential approaches and that the emphasis on evidence hierarchies of methodology fails to lend credence to the common practice of corroboration in medicine. I argue that the strength of evidence lies in the evidence itself, and not the methodology used to obtain that evidence. Ultimately, when it comes to evaluating the effectiveness of medical interventions, it is the evidence obtained from the methodology rather than the methodology that should establish the strength of the evidence.

Keywords: evidence-based medicine, corroboration, meta-analysis, randomized controlled trials, mechanisms, quality of evidence

Introduction

Evaluating evidence is an essential part of clinical medicine. Modern medicine relies on a plethora of scientific methodologies for producing evidence of the sort that can be used reliably for managing treatments. Methods that are currently available for this purpose include randomized trials, observational studies, systematic reviews, and mechanistic reasoning.

The last two decades have seen intensive debates regarding the evidential role that should be ascribed to different methodologies or study designs. Proponents of evidence-based medicine (EBM) have strongly endorsed randomized trials (or, if available, systematic reviews of randomized trials) as the most reliable evidential source for clinical decisions [1]. However, this emphasis on randomized trials has been strongly criticized by several authors, who have questioned the status of randomized trials as the primary source of evidence [2-7]. Meanwhile, other critics have pointed out the seemingly unjustified exclusion of mechanistic reasoning from most hierarchies of evidence [8-11].

Recent developments have led to some important changes in the EBM outlook, for example regarding the evidential role of mechanistic evidence [12-13] and the importance of evidence from non-randomized studies (especially evidence of "large effects") [14]. Still, it appears that the general idea about the epistemic superiority of randomized trials over other types of non-randomized studies persists within EBM. As indicated by the authors of the influential Cochrane Handbook for Systematic Reviews of Interventions, "Potential biases are likely to be greater for non-randomized studies compared with randomized trials, so results should always be interpreted with caution when they are included in reviews and meta-analyses. Particular concerns arise with respect to differences between people in different intervention groups (selection bias)..." [15].

It is curious, however, that there has been no clear evidence to date to support the claim, often ascribed to the proponents of EBM, that randomization reduces selection bias. A number of methodological reviews that have compared the results of randomized studies and non-randomized studies found little evidence that the absence of randomization is associated with selection bias or larger estimates of effect [16-18]. For example, a review conducted by Benson and Hartz in 2000 found "little evidence that estimates of treatment effects in observational studies reported after 1984 are either consistently larger than or qualitatively different from those obtained in randomized, controlled trials" (p. 1878) [19]. In fact, several methodological reviews indicate that high-quality observational studies and randomized trials yield very similar results [20-23].

In contrast to much of the philosophical and empirical work on the assessment of evidence in EBM, this paper argues that considerations regarding the quality of evidence are more important than considerations about choice of study design. My proposition is that the primary concern in testing and managing treatments in clinical practice should be evaluating the strength, quality and relevance of evidence, not merely the nature of the methodology. More specifically, I present the argument that confounding is more likely to be correlated with low quality evidence rather than the nature of the study design. This paper will also suggest that the official EBM view to some extent misrepresents its own practice, in which the results from different methodologies are corroborated together in clinical reasoning, including the much debated reasoning from physiological mechanism [8,9,24].

The next section examines the problem of evaluating the quality of evidence associated with different methodologies. In the following section on strength of evidence, I consider a recent attempt by Howick, Glasziou and Aronson to delineate the evidential role of evidence from a specific methodology (mechanistic reasoning) [13]. I argue that the criteria provided by their accounts fail on two points: circularity and consistency. I present a case for shifting the focus from the justification of methodology to the justification of the evidence itself. The section on corroborating evidence argues that the successful corroboration of evidence for testing treatments and managing disease should aim to limit a priori judgments concerning the relation between evidence and study design as far as possible.

What underwrites confidence in evidential claims?

When it comes to clinical research, the "gold standard" of evidence (i.e., randomized trials) makes it: (a) easier to obtain approval for marketing new drugs [25], (b) more likely studies will be included in a systematic review [15], and (c) more likely the findings will survive critical appraisal [26]. The EBM movement considers the rationale underlying the use of such evidence to be clear: randomization reduces selection bias (i.e., differences between study groups at baseline) [9]. Moreover, randomization is thought to yield more accurate, translatable results than that produced by non-randomized studies, which are usually considered to suffer from potentially powerful biases. For example, the editors of the Cochrane Handbook make the claim: "The logical reason for focusing on randomised controlled trials in Cochrane reviews is that randomisation is the only means of allocation that controls for unknown and unmeasured confounders as well as those that are known and measured" (p. 84) [27].

Based on the assumption that randomization is the standard against which other study designs should be judged, one would expect randomized trials to provide different results from non-randomized studies. Yet, obtaining evidence in support of the "gold standard" has proven to be elusive [16]. A number of meta-analyses comparing randomized trials and observational studies of the same intervention have found no large systematic differences between the two types of studies [28-29]. There have been several reviews indicating that high-quality observational studies and randomized trials yield very similar results [19-23]. To date, there has been considerable disagreement about the question of whether randomization reduces confounding more than non-randomized studies do, and there is no clear evidence to support either position.

Arguments in favour of either position have been presented on logical or theoretical grounds, instead of empirical grounds [30]. Howick [9], for example, has made an interesting case for the epistemic superiority of randomized trials compared with observational studies. He justifies his claim that randomized studies are generally better at estimating treatment effects than non-randomized studies by appealing to the epistemic desideratum "that better evidence rules out more confounders" (p. 52) [9]. This claim apparently points to randomized trials as the superior study design. However, it is curious how one should evaluate the claim that evidence from one methodology is better than evidence from another methodology, especially in the absence of empirical support for the claim that "randomization rules out more confounders". Moreover, even if such evidence is procured, many would still question the current practice of evidence-ranking systems that exclude evidence on purely methodological grounds without regard to evidential quality [30].

It might be more useful, however, to consider the association between confounding and quality of evidence. An interesting example is a meta-epidemiological study by Panagiotou and colleagues that set out to compare treatment effects from randomized trials conducted in more developed countries, to those in less developed countries [31]. Their study indicates that results from less developed countries are more likely to produce favourable estimates of treatment effects for the same treatment intervention. The authors suggest: "Given the systematic preponderance of more favourable results in trials from less developed countries, one potential explanation is that the available randomised evidence from developed countries is more biased" (p. 5) [31]. In other words, "results from low-quality studies are more likely to be biased and therefore tend to overestimate treatment effects". Analyses of possible sources of this bias indicated that the inflation in treatment effects appeared to be due to the lack of quality controls in randomized trials [31].

To be sure, the same phenomenon is likely to hold across study designs. In general, however, the results (effect estimates) from randomized trials are often found to contradict each other. The same holds true for results from observational studies [32]. Many take this as evidence for the lamentable state of evidential quality in existing studies. The poor quality of the available evidence is often remarked upon in methodological reviews, as in this review by Kunz and Oxman [33]:

> As with any review the quality of the data is limited by the quality of the studies that we have reviewed. Most of the studies included in the review had one or more methodological flaws. In many of the included comparisons, particularly those between randomised controlled trials and historically controlled trials, methodological differences other than randomisation may account for some of the observed differences in estimates of effect (p. 1188) [33].

The above statement can be taken to suggest that more effort should be directed towards producing and discerning evidence of a high quality regardless of study type. Clearly, more evidence of high quality needs to be included in systematic reviews and meta-analyses. The effort should, however, extend beyond the question of evaluating study designs. Arguably, there is little reason in, for example, to exclude high-quality observational studies from a meta-analysis on purely methodological grounds. A similar point is made by Shrier and colleagues in an article titled "Should meta-analyses of interventions include observational studies in addition to randomized trials?", where the authors concluded that "...the theoretical and empirical evidence presented in this paper suggests that excluding observational studies in systematic reviews, a priori, is inappropriate

and internally inconsistent with an evidence-based approach" p. (1208) [30].

More focus should be put on quality measures, in light of current EBM guidelines that recommend excluding non-randomized studies (regardless of quality) from meta-analyses. For instance, reviewers from the same Cochrane review group tend to decide to pool the results when comparing randomized trials [27], even if there is considerable heterogeneity (e.g., I^2=87%), as in the Cochrane review on placebos [34]), but they will not do the same in reviews with less heterogeneity that contain both randomized and non-randomized trials (e.g., I^2=77%), as in [35]) [36]. The idea that high-quality evidence from non-randomized studies should not be excluded from meta-analyses on an a priori basis is in sharp contrast to the recommendations found in the Cochrane Handbook (13.2.1.1): "We strongly recommend that review authors should not make any attempt to combine evidence from randomized trials and NRS [non-randomized studies]" [15].

Meanwhile, meta-analyses continue to be conducted on a limited number of randomized trials. As such, the current EBM practice does not necessarily represent the ideal procedure for corroborating evidence. A 2003 study that examined a random sample (1%) of meta-analyses in systematic reviews that were conducted by the Cochrane Collaboration, found that 6 out of 16 reviews included 2 studies or less in their meta-analysis [37].

On a similar note, evidence-ranking systems such as Oxford's "Levels of evidence" [26] and GRADE [38] that judge the evidential quality of study designs, automatically assign lower levels of quality to evidence from non-randomized studies. This type of evaluative procedure is not acceptable when one considers that evidence from studies with the same study design is often found to be contradictory. Generally, one can blame this discrepancy on poorly constructed studies, or one can "bite the bullet" and state what everybody already knows–that the methodology is not perfect. In fact, it is doubtful that any methodology can ever be 100% accurate. There will always be unaccounted for confounding variables. Moreover, it seems less likely that high-quality studies (irrespective of study design) are more contradicting than low-quality studies from the same study design.

Of course, this calls for a more pronounced distinction between, on the one hand, the methodology of choice, and, on the other hand, the evidence obtained from that methodology. A basic assumption here is that even if the methodological choice at hand is faulty or incomplete (which is likely the case for all methodologies), it might yet manage to produce good evidence (as was the case with Ptolemy's methods, which made more accurate predictions than did Copernicus's), or the other way around. This example further suggests the need for an evaluative feature in EBM that is (as far as possible) autonomous from the methodology through which the evidence is obtained.

Before my point is developed in more detail, we will consider a recent argument to incorporate evidence from physiological mechanisms into EBM. Part of the argument entails the quality of the physiological evidence and its corroboration.

Strength of evidence as relative to study design

A large part of the philosophy of science attempts to understand the role of mechanisms in causal inferences and descriptions in the biological and biomedical sciences [39]. Related to this work, several arguments have proposed to include mechanism-based reasoning in EBM [10,24,40]. In this section I will present one of the most influential arguments for incorporating knowledge about physiological mechanisms in EBM and, consequently, show why the proposal fails.

Howick, Glasziou and Aronson [13] have argued persuasively for ways of recognizing high-quality mechanistic reasoning and incorporating it as evidence in EBM. According to their view, mechanistic evidence becomes evidence-based (i.e., of high quality) if two conditions are satisfied. The first condition is that the evidence from mechanistic reasoning is not incomplete [13]. This means that each link in the evidential chain–the input-output relationship–linking the intervention and the outcome is substantiated by further evidence. Further evidence, in this case, means evidence of a different kind, ideally, the kind obtained from randomized trials. The second condition requires taking into account the probabilistic and stochastic nature of mechanisms.

I find the latter condition agreeable since it cautions against oversimplified thinking about mechanistic processes. It is often the case in many mechanistic domains that the mechanisms (e.g., the firing of neurons) are highly probabilistic and stochastic, even if all baseline factors have been taken into account [41]. Leaving the second condition aside, there are two problems with the first condition.

First, taking evidence from randomized trials as support for mechanistic evidence means that the evidence is no longer mechanism-based, but randomized-trials-based. Clearly then, justifying the available mechanistic evidence would not depend on mechanistic reasoning, but on the results from a randomized trial. It is, therefore, circular to claim that the evidence provided by randomized trials are more accurate based on any reported differences between evidence from randomized trials and mechanistic reasoning (or any other non-randomized study methods for that matter).

Comparably, false models or theories can be used to predict favourable outcomes, as illustrated by the case of the Ptolemaic model's superiority over the Copernican model for predicting planetary motions. But this does not mean that the model or theory is correct. It is possible that a randomized trial can find evidence of therapeutic effectiveness following faulty mechanistic reasoning. Many examples can be used to illustrate this point, including the famous sildenafil (Viagra) and atomoxetine (Strattera) trials. The importance of this perspective lies in its exposing the fact that evidence of mechanisms from randomized trials is evidence according to randomized trials, and renders the purported evidential role ascribed to high-quality mechanistic evidence in EBM to be nothing but illusory.

Second, it seems that the general idea of Howick et al. [13] that "knowledge of mechanisms upon which mechanistic reasoning is based, is not incomplete" (p. 438) is somewhat inconsistent. Notice that the employed double negative "not incom-

plete" is a logically equivalent way of saying "complete". Arguably, the conclusion that follows from (1) "reasoning based on empty or partial mechanisms should be disregarded" (p. 439) and (2) "all mechanisms are, at least to some extent, 'partial', in the sense that they are not completely understood" (p. 436) implies that all mechanistic evidence should be disregarded.

These objections point to the more general difficulty of clearly delineating evidence from different methodologies. Moreover, because the methodologies are deeply intertwined, there is often no way of measuring their respective contribution with regard to the evidence obtained. To take mechanistic-reasoning as an example, inferences from mechanisms are frequently found embedded in strategies for testing treatments. Clinical trials rely on evaluations of diagnostic tests for including or excluding patients in studies. Meanwhile, diagnostic tests inevitably rely on functional, physiological or biological evaluations. It is in this specific sense that mechanism-based rationale is an indispensable part of clinical reasoning (i.e., corroboration). In a similar way, answers to questions related to prognosis will tend to rely on the use of biomarkers.

There are cases when biomedical science that is well-characterized with biomarkers or in-vitro diagnostics will gain expedited approval by the biomedical community (such as certain Phase 1 studies of potential breakthrough therapies that were recently approved by the FDA without randomized trials). For instance, in their guideline for clinical evaluation of diagnostic agents the European Medicines Agency (EMA) (an organization analogous to the American FDA) recommends that clinical trials should be designed and conducted with consideration to, amongst other things, the "anatomical condition of interest", "physical attributes", and the "interpretation of imaging results (e.g., inflammation, trauma)" [42]. For example, in evaluating imaging results, the same guideline suggest: "In functional imaging or pathophysiological explorations, the assessment of biological or physiological processes may form the basis for an approval."

Another example to consider is the practice of neuroscience, where neurobiologists, pharmacologists, psychologists and cognitive neuroscientists incorporate numerous evidential justifications for the study of diseases related to the human brain. These methods are integrated across many levels of investigation, ranging from sub-molecular to cognitive and psychological aspects of human behaviour. In addition, the science utilizes several imaging techniques, such as MRI, fMRI, MEG, EROS, CT, PET and others (many of which are used for clinical purposes such as diagnostics examinations). Generally, these scientists do not consider the merits of the evidence they have obtained by primarily asking whether the method used is plausible. Instead, they emphasize other questions, such as: are we justified in believing that the phenomenon discovered exists based on our evidence for its existence? They do not normally ask: are we justified in believing that this phenomenon exists based on the methods we have used for providing evidence for its existence? Relatedly, the management of Alzheimer's disease relies on a similar plethora of methods and imaging techniques. For example, without a proper diagnosis, what reason is there to conduct a randomized trial or do an observational study to see if a treatment works? Diagnosis relies on mechanistic reasoning, as much as randomized trials and observational studies rely on diagnosis.

The present point is that medicine is no exception to other sciences, which normally use a patchwork of methods and models for researching and mapping empirical phenomena [43]. Consider, for instance, the following passage in an official EBM handbook by Straus, Glasziou, Richardson and Haynes [1] on the topic of identifying differences in response to interventions within the same subgroup of patients:

> To summarize [the recommendations], unless the difference in response makes biological sense, was hypothesized before the trial, and has been confirmed in a second, independent trial, we'd suggest that you accept the treatment's overall efficacy as the best starting point for estimating its efficacy in your individual patient (p. 89) [1].

Notice the sheer variety of evidential sources that the authors enumerate: to see if the patient's response makes "biological sense" (mechanistic reasoning), to consider whether the difference in result was anticipated (by experts), and finally, to survey the literature for a different trial that might explain the difference in response (observational studies, randomized or non-randomized trials). This example clearly illustrates the messy, and, at times, seemingly contradictory nature of evidence-gathering practices in medicine.

The recommendation provided by Straus and colleagues [1] and similar examples [44] are suggestive of the common practice of corroborating evidence of treatment efficacy by multiple sources of evidence. As we saw, this relates, for example, to mechanistic evidence obtained from diagnostic and prognostic procedures such as screenings, biopsies and autopsies. Even more straightforward cases as, for example, those involving dramatic effects [14] will also require further corroboration to determine their usefulness in a clinical setting. But the critical element for successful clinical reasoning is obtaining evidence of high quality. The fact that clinical reasoning (i.e., what clinicians do all the time) involves a patchwork of evidential approaches is not mentioned often in the EBM literature.

Corroborating the evidence

Many epidemiologists rightly emphasize the importance of looking at how well a study has been conducted (not necessarily focusing on its design) and how it compares in terms of the quality of similar studies. As indicated, research on the effect of study design on results is yet to be conducted. Indeed, future research might reveal that randomized trials are, on average, better than non-randomized trials or observational studies (see e.g., [45]). However, this would not mean that every individual randomized study is better that every individual observational study.

Typically, different evidence-ranking systems schemes allow for evidence from different study types to be downgraded or

upgraded depending on the quality of evidence (e.g., [26,38]). This option is, however, seldom utilized. Besides, most evidence grading systems are set up in a way that confers a higher initial value to randomized trial study designs, which inevitably upsets the balance at baseline and precludes a fair comparison from the start. This not only makes upgrading difficult, but effectively hinders the possibility of, for instance, combining upgraded non-randomized studies of high quality with randomized studies of similar quality in systematic reviews and meta-analysis. As suggested earlier, the reason for not including high-quality evidence from non-randomized trials is found in the Cochrane Handbook, in the following guideline for reviewing non-randomized studies (NRS): "We strongly recommend that review authors should not make any attempt to combine evidence from randomized trials and NRS" [15]. The problem with this guidance is that it conflicts with the handbook's recommendation that "[m]eta-analysis should only be considered when a group of studies is sufficiently homogeneous in terms of participants, interventions and outcomes" (p. 137) [27]. To say the least, the latter instruction arguably qualifies those non-randomized studies that do fulfil the homogeneity requirement. Therefore, without independent justification, the proviso cannot be relied upon to support a systematic reviewer's decision not to pool randomized and non-randomized studies that show no significant differences in results.

I do not suggest that it is completely unreasonable that some methods can be more relevant for answering certain clinical queries [44]. For example, case controls might be very appropriately used for answering questions related to rare diseases, but they are not ideal for inquiries about short-term intervention effects. However, even in such cases the assessment of evidential quality should be, as far as possible, uninfluenced by theoretical pre-conceptions or a priori judgments relating to the study method of choice. This is so, because the same method can produce contradictory evidence of good quality.

It is seemingly true that meta-analyses found in systematic reviews do, to a certain extent, corroborate evidence by combining evidence from multiple studies. Yet, evidence from high-quality non-randomized studies is not usually corroborated with evidence from randomized trials. This is a problem when one considers that results from investigations of the same interventions conducted by the same methods are not found to be very consistent. Observational studies, when they are not contradicting themselves, often contradict randomized trials. In turn, randomized trials contradict themselves as frequently as they contradict observational studies. Generally, one can blame the latter on poorly constructed studies, or one can acknowledge that the methodology used in randomized trials is not always perfect. In fact, it is doubtful that any methodology can ever be 100% accurate. There will always be unaccountable confounders and other uncontrolled factors. Regrettably, the act of randomizing cannot be taken as a sufficient precept when discerning relevant, reliable research evidence to include in systematic reviews. Ultimately, our best bet is to be as meticulous in evaluating evidential strength as is scientifically feasible. Thus, there should be cross-corroboration of evidence between randomized trials and non-randomised trials, in which high quality evidence from both sources is pooled.

Conclusion

This paper argues for shifting the focus away from the method producing the evidence to the evidence as it stands by itself. The fact that some treatments are effective, while others are not, is separate from the purported quality of the method that ascertains their causal role. I have argued that the same methodology can produce conflicting results. In contrast, high-quality evidence, irrespective of study design, is more likely to be consistent and can be corroborated for the effective treatment and management of disease. In conclusion, I hope to have shown why medicine should be more evidence-based and less method-based.

Acknowledgments

Many thanks to Jeremy Howick for helpful discussions and comments on earlier drafts of this paper. I would also like to thank two anonymous reviewers for their much valued suggestions for improving the manuscript.

References

[1] Straus, S. E., Glasziou, P., Richardson, W. S. & Haynes, R. B. (2011) *Evidence-based Medicine: How to Practice and Teach EBM*, 4th edn. Edinburg: Churchill Livingston.

[2] Cartwright, N. (2007). Are RCTs the gold standard? *Biosocieties, 2*, 11–20.

[3] Russo, F. & Williamson, J. (2011) Epistemic causality and evidence-based medicine. *History and Philosophy of the Life Sciences*, 33, 563–582.

[4] Thompson, R. P. (2010) Causality, mathematical models and statistical association: dismantling evidence-based medicine. *Journal of Evaluation in Clinical Practice*, 16, 267–275.

[5] Worrall, J. (2002) What evidence in evidence-based medicine? *Proceedings of the Philosophy of Science Association*, 3, S316–S330.

[6] Worrall, J. (2007) Evidence in medicine and evidence-based medicine. *Philosophy Compass*, 2, 981–1022.

[7] Worrall, J. (2010) Evidence: philosophy of science meets medicine. *Journal of Evaluation in Clinical Practice*, 16, 356–362.

[8] Andersen, H. (2012) Mechanisms: what are they evidence for in evidence-based medicine? *Journal of Evaluation in Clinical Practice*, 18, 992–999.

[9] Howick, J. (2011) *The Philosophy of Evidence-Based Medicine*. Chichester: Wiley Blackwell & BMJ Books.

[10] Russo, F. & Williamson, J. (2007) Interpreting causality in the health sciences. *International Studies in Philosophy of Science*, 21, 157–170.

[11] Bluhm, R. (2007) Clinical trials as nomological machines: implications for evidence-based medicine. In *Establishing Medical Reality* (eds H. Kincaud & J. McKitrick), pp. 149–166. New York: Springer.

[12] Sackett, D. L. (1996) Evidence based medicine. What it is and what it isn't. *British Medical Journal*, 312, 71–72.

[13] Howick, J., Glasziou, P. & Aronson, J.K. (2010) Evidence-based mechanistic reasoning. *Journal of the Royal Society of Medicine*, 103:433–441.

[14] Glasziou, P., Chalmers, I., Rawlins, M. & McCulloch, P. (2007) When are randomised trials unnecessary? Picking signal from noise. *British Medical Journal*, 334, 349–351.

[15] Higgins, J.P.T. & Green, S. (editors). (2011) *Cochrane Handbook for Systematic Reviews of Interventions Version*. 5.1.0 [updated March 2011]. The Cochrane Collaboration. Available at: www.cochrane-handbook.org.

[16] Vandenbroucke J.P. (2011) Why do results of randomized and observational studies differ? *British Medical Journal*, 343. d7020. DOI: 10.1136/bmj.d7020

[17] Bero, L., Anglemyer, A. & Horvath, T. (2012) Healthcare outcomes assessed with non-experimental designs compared with those assessed in randomised trials.*Cochrane Database of Systematic Reviews*, February 15 (2): MR000034. DOI: 10.1002/14651858.MR000034.

[18] Ioannidis, J., Haidich, A.B., Pappa, M., Pantazis, N., Kokori, S.I., Tektonidou, M.G., et al. (2001) Comparison of evidence of treatment effects in randomized and non randomized studies. *Journal of the American Medical Association*, 286, 821–830.

[19] Benson, K. & Hartz, A.J. (2000) A comparison of observational studies and randomized, controlled trials. *New England Journal of Medicine*, 342, 1878–1886.

[20] Britton. A., McKee, M., Black, N., McPherson, K., Sanderson, C. & Bain, C. (1998) Choosing between randomised and nonrandomized studies: a systematic review. *Health Technology Assessment*, 2(13), 1–124.

[21] Concato, J., Shah, N. & Horwitz, R.I. (2000) Randomized controlled trials, observational studies, and the hierarchy of research designs. *New England Journal of Medicine*, 342, 1887–1892.

[22] MacLehose, R.R., Reeves, B.C., Harvey, I.M., Sheldon, T.A., Russell, I.T. & Black, A.M. (2000) A systematic review of comparisons of effect sizes derived from randomised and non-randomised studies. *Health Technology Assessment*, 4(34), 1–154.

[23] Oliver, S., Bagnall, A.M., Thomas, J., Shepherd, J., Sowden, A., White, I., et al. (2010) Randomised controlled trials for policy interventions: a review of reviews and meta-regression. *Health Technology Assessment*, 14(16), 1–165.

[24] Bluhm, R. (2013) Physiological mechanisms and epidemiological research. *Journal of Evaluation in Clinical Practice*, 19, 422–426.

[25] FDA (2005) Part 314: *Applications for FDA Approval to Market a New Drug*. United States Food and Drug Administration, Silver Spring, MD.

[26] *Levels of Evidence*. Oxford Centre for Evidence-Based Medicine (2011). Available at: http://www.cebm.net/index.aspx?o=5653. (Accessed 10 Jan 2014).

[27] Higgins, J.P.T. & Green, S. (2006) (editors) Cochrane Handbook for Systematic Reviews of Interventions 4.2.6. The Cochrane Library, Issue 4. Chichester, UK: John Wiley & Sons.

[28] Ioannidis, J. & Lau, J. (2001) Evolution of treatment effects over time: empirical insight from recursive cumulative metaanalyses. *Proceedings of the National Academy of Sciences of the United States of America*, 2001, 98(3), 831–836.

[29] Furlan, A.D., Tomlinson, G., Jadad, A., et al. (2005) *Why randomized trials and non-randomized studies of the same interventions agree or disagree*. Paper presented at the 4th Canadian Cochrane Symposium, Montreal, Quebec, Canada, March 12, 2005.

[30] Shrier, I., Boivin, J.F., Steele, R.J., Platt, R.W., Furlan, A., Kakuma. R., Brophy, J. & Rossignol, M. (2007) Should meta-analyses of interventions include observational studies in addition to randomized controlled trials? A critical examination of underlying principles. *American Journal of Epidemiology*, 166, 1203–1209.

[31] Panagiotou, O.A. & Contopoulos-Ioannidis, D. G., Ioannidis, J.P.A. & Rehnborg, C.F. (2013) Comparative effect sizes in randomised trials from less developed and more developed countries: meta-epidemiological assessment. British Medical Journal, 346. f707. DOI: 10.1136/bmj

[32] Ioannidis, J.P. (2005) Contradicted and initially stronger effects in highly cited clinical research. *Journal of the American Medical Association*, 294(2), 218–228.

[33] Kunz, R. & Oxman, A.D. (1998) The unpredictability paradox: review of empirical comparisons of randomised and nonrandomised clinical trials. *British Medical Journal*, 317, 1185–1190.

[34] Hróbjartsson, A, & Gøtzsche, P.C. (2010) Placebo interventions for all clinical conditions. *Cochrane Database of Systematic Reviews*. January 20, (1):CD003974. DOI: 10.1002/14651858.CD003974.pub3.

[35] Odgaard-Jensen, J., Vist, G.E., Timmer, A., Kunz, R., Akl, E.A., Schnemann, H., et al. (2011) Randomization to protect against selection bias in healthcare trials. *Cochrane Database of Systematic Reviews*, April 13, (4): MR000012. DOI: 10.1002/14651858.MR000012.pub3.

[36] Howick J. & Mebius A. (2014) Unpredictable evidence that randomization protects against bias in healthcare trials. Manuscript submitted for publication.

[37] Shrier I. (2003) Cochrane Reviews: new blocks on the kids. *British Journal of Sports Medicine*, 473–474.

[38] Guyatt, G.H., Oxman, A.D., Kunz, R., et al. (2008) Rating quality of evidence and strength of recommendations: What is "quality of evidence" and why is it important to clinicians? *British Medical Journal*, 336, 995–998.

[39] Mebius, A (2014) A weakened mechanism is still a mechanism: On the causal role of absences in mechanistic explanation. *Studies in History and Philosophy of Biological and Biomedical Sciences*, 45, 43–48.

[40] Howick, J. (2012) Exposing the vanities–and a qualified defense–of mechanistic reasoning in health care decision making. *Philosophy of Science*, 78, 926–940.

[41] Craver, C. F. (2007) *Explaining the Brain: Mechanisms and the Mosaic Unity of Neuroscience*. New York: Oxford University Press.

[42] European Medicines Agency Committee for medical products for human use (CHMP) (2009). Guideline on clinical evaluation of diagnostic agents. Available at www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2009/09/WC500003580.pdf. (accessed 21 December 2013).

[43] Cartwright, N. (1999) *The Dappled World: A Study of the Boundaries of Science*. Cambridge: Cambridge University Press.

[44] Chalmers, I., Glasziou, P. & Vandenbroucke, J.P. (2004) Assessing the quality of research. *British Medical Journal*, 328, 39–41.

[45] Savovic, J., Jones, H.E., Altman, D.G. et al. (2012). Influence of reported study design characteristics on intervention effect estimates from randomized, controlled trials. Annals of Internal Medicine, 157, 429-438.