The Journal of Medical Research

Flattening the Truth Pyramid: Reconsidering the Evidence Hierarchy

Steve Bierman[1]

[1] *University of Arizona, College of Medicine. Tucson, Arizona. USA*

Abstract

For decades, the evidence hierarchy - which places randomized controlled trials (RCTs) and systematic reviews at its apex – has contributed positively to decision-to-treat considerations. Nonetheless, RCTs have often failed to reveal efficacy and safety concerns relating to the studied treatments. Moreover, many of the questions posed in clinical practice are best answered by means other than RCTs. This discussion examines the flaws in RCT statistical methodology that contribute to their limitations. It also presents novel methodologies, that combine RCT data with observational data, and thereby enable clinicians to make personalized treatment decisions for individual patients – something RCTs alone cannot do. Finally, this discussion explores what constitutes the best evidence to answer the many questions clinicians confront on a daily basis. The upshot is a flattened evidence hierarchy wherein RCTs, observational studies and novel methodologies are placed in their proper context, so that their relevance to clinical medicine is neither exaggerated nor ignored.

Keywords: Evidence-based medicine, Randomized controlled trials, EBM, RCT, Clinical trial, Reproducibility, Precision medicine, Causal inference, Statistical flaws.

INTRODUCTION

Under the aegis of Evidence Based Medicine (EBM) randomized control trials (RCTs) and their systematic reviews have been elevated to the apex of the truth pyramid, while anecdotes and other observational information have been relegated to the netherworld of suspicion and doubt [1-4]. This evidence hierarchy, as it is called, was initially promoted by Shannaussy and others to encourage refinements in the decision-to-treat methodologies of practicing clinicians [5]. However, owing in part to its adoption by Cochran and other investigative and regulatory institutions, the evidence hierarchy is now afforded an almost-scriptural status. The result, in both medical research and clinical practice, is a reflexive disdain for all things not-RCT – namely, anecdotes, case series and observational studies [6].

Notwithstanding their ascendency, RCTs have failed to produce the relevant efficacy and safety data their heightened status would seem to imply. With respect to efficacy, the top ten drugs sold in the US – all studied in RCTs - fail to improve the condition being treated in 75-96% of patients who take them [7]. With respect to safety, prescription medicines have now become the third leading cause of death in the US and Europe [8].

The first purpose of this paper is to demonstrate the inherent limitations of RCT methodology as applied in clinical medicine [9]. The second purpose is to elevate observational data to its rightful place, through consideration of new and exciting developments from the world of causal inference [10]. Finally, we intend to place the RCT in its proper context, so that its relevance to clinical medicine is neither exaggerated nor ignored.

LOOKING MORE CLOSELY AT RCTs

Even in the absence of bias and influence there are fundamental limitations to RCTs that require our present attention. Understanding these limitations will lend caution to our interpretation and usage of RCT results [9].

COMPARISON OF AVERAGES

"The core methodological idea in clinical trials is the comparison of averages." (ibid, pg. 4) However, this methodology can be and often is misleading. As Hanin observes: "The seemingly appealing idea that the

*Corresponding author:
Dr. Steve Bierman
University of Arizona, College of Medicine. Tucson, Arizona. USA
Email: stevebierman00@gmail.com

best intervention is the one that works best on the average may be true in the case of homogeneous responses. However, as a general comparison principle, it represents a fundamental fallacy." (ibid, pg.4)

Consider this scenario. Two treatments (#1, #2) are each compared with the standard treatment (#3). All sample populations are matched for age, gender, underlying conditions, medications and other typical characteristics. The magnitude of the treatment responses are as follows:

Table 1: Magnitude of the treatment responses in populations

Treatment	One Half of Sample	Other Half of Sample	Average Outcome
Treatment #1	0	2	1
Treatment #2	2	0	1
Treatment #3	1.5	1.5	1.5

Standard statistical comparison of averages for this data will show treatment #3 to be the most efficacious. Accordingly, treatments #1 and #2 will be dismissed as inferior. However, we should note that treatment #3 is, in fact, 25% less effective than the other treatments on the responsive halves of their respective sample populations. The statistical comparison of averages, has failed to reveal the very real - albeit, circumscribed – efficacy of treatments #1 and #2.

Moreover, we should note that this comparison takes into account only benefits. Of course, RCTs do record side effects. But again, a comparison of averages with respect to adverse events – especially when the data from various substrata are aggregated - would not pinpoint which individuals or strata are most prone to adverse events.

SAMPLE HOMOGENEITY

Another fundamental assumption underlying the statistical analysis of RCT data is sample homogeneity. However, the data above demonstrate that sample homogeneity is not the case in the example cited, despite the samples being "well-matched." As is often the case, our parameters for matching RCT participants have somehow failed to achieve homogeneity within and between sample populations with respect to relevant parameters.

Imagine, for example, that two traits (A and B) are discovered after this clinical trial is completed. Such traits might be invisible (e.g. a serum biomarker, a mental characteristic) or visible (e.g. some phenotypic characteristic). The responsive half of Sample #1 is positive for trait A, while the unresponsive half is negative for trait A. Both treatment #1 halves are negative for trait B.

Similarly, the responsive half of Sample #2 is positive for trait B, while the unresponsive half is negative for trait B. Again, both treatment #2 halves are negative for trait A.

Further, we discover that all of the treatment #3 sample population is positive for both traits A and B.

Now, all we need do is screen for traits A and B in order to decide which treatment will be most effective for any individual. Those positive for trait A and negative for B will receive treatment #1. Those positive for trait B and negative for A will receive treatment #2. And, finally, those positive for both A and B will receive treatment #3.

What should concern us here is that when treatment #1 or #2 are chosen, a clinical response will ensue that is superior to treatment #3 – the standard treatment, and the one favored by the RCT.

There is no reason to expect this hypothetical situation does not obtain in many, if not most, RCTs conducted today. For example, Hanin demonstrates that sample heterogeneity with respect to outcome potential has seriously hampered contemporary breast cancer treatment research for decades. (ibid, pg. 4)

IDOLATRY OF p-VALUES

Then, there is the idolatry of *p-value*s which today persists in medical science despite its growing disfavor in other fields. Using $p \leq 0.05$ as the threshold for rejecting the null hypothesis, was first proposed by Ronald Fisher in a low-key manner in his landmark book [11]. It has since become an almost religious commandment for medical researchers. But the selection of this arbitrary threshold is a flawed method for biomedical research, largely because: (1) it entails an underlying distribution under the null hypothesis that is unrealistic, and (2) it relies on the tacit assumption of a one- or two-sided bell-shaped tail, which most clinical trials can, at best, only approximate. "Therefore," as Hanin concludes, "for sample sizes typically encountered in clinical trials…the maximum error in *p-value* determination may be comparable to, or even exceed, the small *p-value*s used for rejecting the null hypothesis. Such sample sizes can only guarantee the correctness of the *first* decimal digit of the *p-value*. Thus, pursuit of small *p-value*s in parametric analysis of clinical trials is indefensible [9]."

SAMPLE SPECIFICATIONS

Finally, consider that most statistical analyses of trial data assume a fixed sample size. Yet, in reality, sample size is rarely fixed; rather, it is most often variable owing to inclusion criteria, exclusion criteria and loss of subjects due to lack of benefit or side effects. As Hanin notes: "Statistical methods intended for fixed sample size lead to erroneous results if applied to samples of random size." (ibid, pg.8) Clearly, this is an important consideration rarely discussed when assessing the validity of RCT results.

SUMMARY

In other words, there is nothing in the application of statistical methodology to RCTs that guarantees the real-world accuracy of its conclusions. In fact, there is much in the messy reality of clinical research that violates the fundamental assumptions of our statistical methods and, therefore, should engage our skepticism concerning the derived conclusions. The fallacy of comparing averages, heterogeneity of sample populations, arbitrariness and inherent flaws of *p-value* thresholds, and the violation of fixed sample size assumptions – these are but a few of the very real concerns every physician should entertain when considering whether and how much to credit statistical conclusions from RCTs. Moreover, these are the very reasons – in addition to internal bias and external influence - that statistical inferences from RCTs are so often false and/or irreproducible.

What about systematic reviews and meta-analyses of RCTs?

To quote a recent article in The Conversation: "A systematic review is only as good as the rigor it employs in combining similar studies of similar interventions with similar measurement of outcomes. When very different studies of different interventions are combined, the results are not informative [12]."

It should be clear that none of the aforementioned concerns regarding RCTs disappear by comparing or pooling data from disparate samples and non-identical studies. Nor does such a review guarantee the elimination of internal biases or external influences. Moreover, data pooling does nothing to solve the fundamental problem clinicians face with RCTs, namely, the inability to extrapolate from the sample population to the individual patient.

Hence, to accord systematic RCT reviews and meta-analyses the uppermost echelon in the evidence hierarchy is to presume that systematic data comparison or aggregation somehow cleanses RCTs of all their inherent flaws, biases and influences. Only in the rarest of cases, when the rigor of the comparison is impeccable, is this true. More commonly, as Judea Pearl recently tweeted of RCT systematic reviews, "It is comparing apples to oranges in hope of finding out something about bananas."

The combination of RCT Data with Observational Data

Fortunately, the work of Scott Mueller and Judea Pearl now points to a solution (or, at least, a partial solution) of the extrapolation issue. Specifically, they demonstrate that the combination of experimental data (e.g. from RCTs) with observational data enables us to do something RCTs and systematic RCT reviews cannot do – namely, define "informative bounds" as to the benefit (or harm) of treatment for specific individuals [13].

Consider, for example, that an independent survey is conducted on the same sample population as described earlier (Table 1). The survey simply asked participants whether they wished to undergo the offered treatment or not.

For the sake of clarity, let us imagine an extreme case. Say, for example, that in the Treatment #1 group all of those patients who did not wish to receive the treatment (but did anyway) experienced zero effect; whereas, all of those patients who did wish to receive the treatment benefitted at a level 2 response rate. And likewise, in Treatment #2, all of those who did not wish to undergo treatment (but, nevertheless, did) experienced zero effect; whereas, all of those patients who did wish to receive treatment also benefitted at a level 2 response rate.

Let us further assume that the survey of Treatment #3 patients revealed the same treatment effect regardless of preference.

Hence, in this hypothetical instance, choice matters in determining the outcomes of Treatments #1 and #2. This is not evident from the RCT data alone, but emerges clearly as a result of combining RCT and observational data.

While this is an extreme example with 100% benefit among the treatment-choosers and 0% among the non-choosers, one could easily imagine survey results of less than 100% for each group. In such a case, Mueller and Pearl teach certain probability calculations that can be used to establish upper and lower bounds on individual causal effects. "These bounds," assert Mueller and Pearl, "… sometimes can be quite narrow and allow us to make accurate personalized decisions." (ibid, pg. 13) Something quite impossible with RCT data alone.

Some might argue that the whims of the patient have no place in modern medicine, that medical outcomes are purely a consequence of biochemical systems, and that this exercise is therefore nothing more than a fabulous stunt. But, as Pearl points out, "the observational data incorporates individuals' whims, and whims are proxies for hidden factors that may affect that individual's response to treatment." He further notes that while confounding factors like individual whims are "usually problematic in causal inference…here confounding helps us, exposing the underlying mechanisms its associated whims and desires are a proxy for." (ibid, pg. 10)

Another application of this approach, whereby RCT results combine with observational data to yield more individualized information, is to calculate the probability of harm from a given treatment for a given individual or class of individuals. This requires the use of notation and mathematics that are generally unfamiliar to physicians. Eventually, one may expect that these equations will be incorporated into software allowing simple inputting of data to produce comprehensible answers for caregivers. Those readers wishing to engage more actively in the mathematics behind the results of this section should refer to Mueller and Pearl [13]. For now, let me summarize the results of this novel approach.

Consider, the following scenario from Mueller and Pearl (Tables 2 and 3):

Table 2: Female Survival and Recovery Data

Experimental	Survival	Death	Total
Drug	489 (49%)	511 (51%)	1,000 (50%)
No Drug	210 (21%)	790 (79%)	1,000 (50%)
CATE*	28% (Net benefit from the Drug)		
Observational	Survival	Death	Total
Drug	378 (27%)	1,022 (73%)	1,400 (70%)
No Drug	420 (70%)	180 (30%)	600 (30%)

*Conditional Average Treatment Effect

Table 3: Male Survival and Recovery Data

Experimental	Survival	Death	Total
Drug	490 (49%)	510 (51%)	1,000 (50%)
No Drug	210 (21%)	790 (79%)	1,000 (50%)
CATE	28% (Net benefit from the Drug)		
Observational	Survival	Death	Total
Drug	980 (70%)	420 (30%)	1,400 (70%)
No Drug	420 (70%)	180 (30%)	600 (30%)

In this scenario, an RCT (Experimental data) was conducted to assess the efficacy of the drug for the treatment of a certain cancer. Notice, the average net benefit from taking the drug versus not taking the drug is 28% for both males and females. Based on these findings (which were statistically significant), a reasonable conclusion might be that for some patients there is real benefit from the drug and, therefore, all patients - male and female - should take the drug in order to increase their chances of survival.

Subsequent to the RCT, to continue this scenario, an independent research body decided to conduct a survey (observational data) to determine how many patients actually took the drug after it was recommended by their doctor. It turns out that only 70% of patients, male and female, actually took the drug after it was recommended. Rumors of adverse side effects and unexpected death apparently dissuaded the other 30% from complying with their doctors' suggestions.

Now, using the method described by Mueller and Pearl which combines the RCT and observational data, we can quantitatively assess the probability of harm from the drug. Notice, the RCT does not allow us to make this determination. Moreover, it suggests an equivalent chance of benefit (28%) in both males and females.

Accordingly, the probability of harm, P(harm) = P(benefit) − CATE: where P(harm) denotes the probability of harm; P(benefit) denotes the probability of benefit, as computed using the method of Pearl and Tian [14]; and CATE denotes the conditional average treatment effect as determined by the RCT (i.e., 28%). In order to avoid more complex calculations, let us accept Pearl and Mueller's calculations of the P(benefit) [13]. Hence, for males P(benefit) = 0.49 and for females P(benefit) = 0.28. Using these numbers, which depend for their derivation on both RCT and observational data, we can now calculate the P(harm) of the drug.

For females, we see that P(harm) = 0.28 − 0.28 = 0. In other words, for females the drug offers the possibility of benefit (28%) without harm. (This is a somewhat counter-inductive conclusion, since looking at the observation study one might have concluded that the risk of death from the drug is greater in women than men. It is not. It turns out that women with more advanced disease took the drug, while women with less advanced disease chose not to take the drug.)

For males, P(harm) = 0.49 − 0.28 = 0.21. Here we discover the emergence of new information that could dramatically affect the shared treatment decisions of doctors and patients. For males, while

the drug offers a 49% chance of benefit, this comes with a 21% chance of harm.

None of this information is accessible from RCT data alone. Yet, as Mueller and Pearl write, "an observational study, however sloppy and uncontrolled, provides a deeper perspective on the treatment's effectiveness. It incorporates individuals' whims and desires that govern behavior under free-choice settings. And since such whims and desires are often proxies for factors that also affect outcomes and treatments (i.e., confounders), we gain additional insight hidden by RCTs." (ibid, pg. 9)

SUMMARY

New methods now allow us to uncover personalized information by combining data from disparate sources – RCTs and observational studies. Alone, neither source is sufficient and both are deeply flawed. However, together they can yield what clinicians have always needed – namely, a solution (or, at least, a partial solution) to the extrapolation problem, whereby the decision to treat becomes less probabilistic and more relevant to each patient.

The Questions Clinicians Ask

So far our attention has been focused on methodologies for determining the best treatment for a given condition. Our question has been: Is one intervention superior to another intervention, or, to no intervention at all? But this is not the only question clinicians must routinely ask and answer; nor is it even the most important one. Indeed - What is this patient's overall health status? What is their diagnosis? What are the causes of this condition? How shall I address these causes? What is this patient's prognosis? – these questions, in addition to treatment decisions - are imbedded in virtually all patient/clinician encounters. As we shall see, each requires a different methodology or combination of methodologies to arrive at an answer.

Let's consider each of the aforementioned questions and ask: What kind of evidence will provide the most useful answers?

WHAT IS THE CURRENT STATE OF THIS INDIVIDUAL'S HEALTH?

An important task for any clinician is determining their patient's state of health - i.e., their recuperative capacity. This is an essential part of every patient's initial assessment. It will influence the clinician's thinking about diagnostic possibilities, treatment options, prognosis and prevention.

It is in this clinical arena that RCTs are, perhaps, least useful. Rather, an individual case study of the patient assumes paramount importance. In this regard, it is worth remembering the words of Hanin: "The great advantage of individual case studies is that learning everything that is there to know is quite feasible and that, in this case, inference from the generated data is not confounded by inter-subject variation. If one believes that biomedical processes are governed by natural laws and have causes, mechanisms and effects, then studying a single subject thoroughly should be very informative." [9]

In assessing an individual's health, we turn to the history, physical examination, laboratory tests, imaging studies, etc. Our assessment will be both qualitative and quantitative. For example, we will want to know the patient's family history of health and illness, and the patient's personal history of illness or injury and recovery. We will also need numbers: Blood counts, electrolyte levels, body weight, etc. And our imaging studies will be useful as well, in both ruling in and ruling out certain maladies.

True, the case series of patients we witness during our own careers will influence us, just as will the case series upon which laboratory "normal ranges" are founded. And RCTs will play a role, too, as they determine at times what tests we order or forego, what questions we ask, what we consider relevant and what we chose to ignore. In fact, all of our compounded knowledge will be brought to bear on this business of assessing a patient's health status.

But at the end of the day, it is the individual case study of the patient that tops the evidence hierarchy in this area of investigation.

WHAT IS THE PATIENT'S DIAGNOSIS?

What evidence is required to diagnose a patient's illness?

Of course, we first need the results of the individual's case study to define his or her signs and symptoms. We then need to know the array of diagnostic categories, the differential diagnosis, into which those signs and symptoms might fall. Generally, this knowledge is derived from descriptive observations, individual case studies and case series. Less commonly, diagnostic categories arise from more sophisticated clinical trials.

Further, we benefit from certain heuristic principles by which to organize our thoughts – most notably Ockham's razor, which combines with deductive reasoning in our attempts to subsume all signs and symptoms under a single diagnostic category. All of this, of course, is buttressed by our broader knowledge of medical science – i.e., anatomy, pathophysiology, genetics, etc. – which supports our ultimate diagnosis.

As you can see, there is no primacy of RCTs in this mix. Rather there is the forever-business of science, grasping for better and better explanations to encompass a wider and wider array of observations and discoveries. Sometimes this quest involves RCTs; other times, it involves serial or individual observations, hypotheses and even trial and error.

WHAT IS THE CAUSE (OR CAUSES) OF THE PATIENT'S ILLNESS?

Traditional allopathic medicine is generally content to ascribe a single cause to a patient's illness. Hence, a patient's pneumonia is caused by the proliferation of pneumococcus in the lungs; a patient's leukemia is the consequence of a genetic mutation; warts are the result of HPV infection; and so on. At times, there are vague references to "triggers" or "stress" which somehow contribute or predispose. But generally, disease is seen as emanating from a single source which ignites a sequence of linear fuses, each lighting the next and the next, until the disease-bomb finally explodes.

What is the evidence for these blunt beginnings - these isolated material causes that purportedly result in our patients' illnesses? As it turns out, the single-source hypothesis of traditional western medicine is presumed, not proven. There is no real evidence behind it, other than that the elimination of the presumed single cause often ameliorates the disease. But removing the intervening billiard green prevents the cue ball from striking the target ball; and yet, the billiard green is never assumed to be the cause of the target ball's motion.

In other words, one of the foundational pillars of modern medicine stands but weakly supported – a bare possibility among many other equally plausible possibilities.

Consider this: traditional western medicine has declined to investigate its single-cause presumption by asking the obvious question, "Why today? Why now?" Pneumococcus, for example, has been a patient's cordial cohabitant for some time. Then, suddenly, it busts out in a rage against the tissues of its once-hospitable host. Why now? Or, the HPV virus has been a ubiquitous neighbor for decades. Suddenly it discovers an unwonted receptivity, and tumors (e.g., warts) form. Why today?

It turns out, when our causal presumptions are probed deeper, a full array of other contributory causes emerge – including contributory noetic causes. When asked why today, for example, the heart attack victim will reply: "I suppose it has something to do with my mother-in-law moving in for the next YEAR!" Or, the appendicitis victim will say, "I

just really needed a rest from work." Or, the breast cancer victim will disclose, "I just feel so guilty for having this ongoing affair."

Of course, these discoveries are only now being described in small case series [15]. Larger case series with well-thought parameters and metrics will surely follow in due course. The point, for our purposes, is to recognize that the multi-cause hypothesis presently exists in opposition to the single cause hypothesis, and that the ascendency of one notion over the other will likely result not only from clever RCTs, but also from the same observational and deductive processes typical in the conceptual evolution of the hard sciences.

HOW SHALL THE PATIENT'S ILLNESSES BE TREATED?

As we have discussed, it is in answer to this questions that RCTs and their systematic reviews can be of most value. They offer us statistical information pertinent to groups of patients. But no RCT or systematic review, in and of itself, can offer us precise information bearing on the patient who sits before us.

Fortunately, the work of Mueller and Pearl offers hope in this connection [10]. For, as we have seen, the combination of observational data and experimental data using the methods of causal inference can produce patient-specific probability bounds, and sometimes even precise personalized predictions.

But in truth, this kind of information addresses only a small fraction of the treatment decisions a practitioner commonly requires. Consider, for example, the decision to suture a fresh laceration. While there may be an RCT that showed the superiority of early wound closure over closure "by secondary intention," for certain kinds of wounds, depending on their depth and location, that RCT may not pertain. Likewise, when the diagnosis is uncertain, the decision to administer a drug as a clinical trial or watchfully await further developments is most often based on prior experience or the teachings of an authority figure.

In other words, when well-designed and well-executed RCTs exist and apply to the circumstances, evidence-based medicine (especially when coupled with methods of causal inference) should prevail. But for the multitude of maladies and conditions that remain unstudied by such means, the teachings of an authority figure, individual observations, case studies and simple logic – all combine to influence treatment decisions.

HOW CAN WE PREDICT A PATIENT'S OUTCOME?

In a word, we cannot predict any one patient's outcome with certainty. We have seen how RCTs fail to justify an extrapolation from group data to individual outcomes - from broad probabilistic information to precision personal information. The fallacy of extrapolation is real; it arises from the enormous individual variability between one individual and the next.

True, the methods of Mueller and Pearl hold promise in allowing for more precise personal predictions [10]. But, except in the rarest instance, these new methods of inference only narrow the bounds of probability; they do not pin a perfect prediction on any patient.

Moreover, the only case study that can guarantee an accurate prediction is one that "matches" the patient perfectly. And that virtually never occurs. For example, if one identical twin develops Parkinson's Disease, the likelihood that the other will develop Parkinson's Disease is not 100%. Rather, it is 5% [16]. Why? Because, even with identical twins, there is no perfect match across all parameters.

And yet often patients do not ask for or need a precise prognosis, especially when the diagnosis is dire. What they ask for is hope. They need to know that their condition is survivable – that they have a chance. Here, an anecdote – a single case report - proves the possibility of survival and thereby ignites the flames of hope. For many patients

and practitioners there is no datum more important. True, a case series can add to the illumination. And, where treatment options exist, an RCT can provide guidance as to which option may be best, and so, further stoke the flames of hope. But it is often the anecdote that holds the high ground in matters of prognostication – at least, from the patient's perspective.

CONCLUSION

Careful analysis of the evidence required to function optimally in the clinical world reveals that the current evidence hierarchy, the so-called truth pyramid, is insupportable. The fact is, all of our evidence sources are flawed. At times, that best evidence will be RCTs or their systematic review. At other times, observational data will best answer our needs and our patients' needs. At still other times, it will be the combination of RCTs and observational studies that provide the best evidence for clinical decision making.

In 2007, Wayne Jonas proposed that the evidence pyramid be flattened and replaced, as he put it, with an "evidence house" [17]. If we envision that house as a single-story, ranch-style structure, the present author agrees. And, to extend the metaphor, in each room a separate clinically relevant question is asked. Sometimes one methodology occupies a room to provide the best answer possible. Other times, two or more methodologies are necessary to enhance the quality and applicability of the answer.

It is time to flatten the truth pyramid - and to accord both RCTs and descriptive evidence their proper place in the evidence house of clinical medicine.

Declarations

1. Ethics approval and consent to participate: Not applicable.

2. Consent for publication: Not applicable.

3. Availability of data and materials: Data sharing is not applicable to this article as no datasets, other than hypothetical sets, were generated or analyzed during the current study.

4. Conflicts of Interest: The author declares he has no competing interests.

5. Funding: Not applicable.

6. Author's contribution: SB is the sole author of this article and approves it for submission; he assumes sole responsibility for the article in its entirety.

REFERENCES

1. Hill N, Frappier-Davignon L, Morrison B. The periodic health examination. Can Med Assoc J. 1979 Nov 3;121:1193-254.
2. Sackett DL. Rules of evidence and clinical recommendations on the use of antithrombotic agents. Chest. 1989 Feb 1;95(2):2S-4S.
3. Burns PB, Rohrich RJ, Chung KC. The levels of evidence and their role in evidence-based medicine. Plastic and reconstructive surgery. 2011 Jul;128(1):305.
4. Centre for Evidence Based Medicine. [Accessed December 17, 2010]; Available at http://www.cebm.net.
5. KNOWING WO. Clinical jazz: harmonizing clinical experience and evidence-based medicine. The Journal of family practice. 1998 Dec;47(6):425-28.
6. Loannidas JP. Hijacked evidence-based medicine: stay the course and throw the pirates overboard. J Clin Epidemiology 2017 April; 84:11-13.
7. Schork N. Time for one-patient trials. Nature 2015;520:609-11.
8. Goetsche P. Deadly medicines and organized crime: how big Pharma has corrupted healthcare. London: Radcliffe Publishing; 2013.

9. Hanin L. Why statistical inference from clinical trials is likely to generate false and irreproducible results. BMC medical research methodology. 2017 Dec;17(1):1-2.

10. Mueller S, Pearl J. Personalized decision making–A conceptual introduction. Journal of Causal Inference. 2023 Apr 19;11(1):20220050.

11. Fisher RA. Statistical methods for research workers. InBreakthroughs in statistics: Methodology and distribution 1970 (pp. 66-70). New York, NY: Springer New York.

12. MacIntyre CR, Chughtai AA, Fisman D, Greenhalgh T. "Yes, masks reduce the risk of spreading COVID, despite a review saying they don't. The Conversation, February. 2023;6.

13. Mueller S, Pearl J. Personalized decision making–A conceptual introduction. Journal of Causal Inference. 2023 Apr 19;11(1):20220050.

14. Tian J, Pearl J. Probabilities of causation: Bounds and identification. Annals of Mathematics and Artificial Intelligence. 2000 Oct;28(1-4):287-313.

15. Bierman SF. HEALING – Beyond Pills & Potions. San Diego: Gyro Press Int'l; 2020.

16. Pelletier KR. Change Your Genes, Change Your Life. Odyssey Press; 2023.

17. Jonas WB. The evidence house: how to build an inclusive base for complementary medicine. Western Journal of Medicine. 2001 Aug;175(2):79.