Perspective | Perspective | Check for updates

Science between evidence and illusion

Wolfgang Lederer*[iD]

Department of Anesthesiology and Critical Care Medicine, Medical University of Innsbruck, Innsbruck, Austria.

*Correspondence:wolfgang.lederer@i-med.ac.at

Abstract

The ever-increasing number of scientific publications lets the half-life of evidence appear shorter and shorter. In scientific research, evidence does not enjoy 100% certainty. Considering the variety of approaches, points of view and constructions with regard to the comprehensibility and perceptibility of evidence, systematic error in scientific studies is immanent. In general, evidence should be well-supported by repeated implementation studies rather than by a single falsified hypothesis in a stand-alone study. Ideally, the estimated probability of certainty should be specified regarding the accuracy of observations, measurements and calculations, and conclusions. A division into categories of certainty may be useful, ranging from obvious (observational studies) and proven (interventional studies) to evident (meta-analyses and systematic reviews). The term evidence as currently used in scientific reporting might provoke higher expectations regarding levels of certainty than are justified. Evidence and bias have to be seen in a complementary context.

Keywords: bias; data science; evidence based medicine; implementation science; philosophy, medical; research design; scientific experimental error

Evidence in Context

• Evidence concerns reality as seen from the momentary structure-centered view. • Evidence and bias have to be seen in a complementary context. • The perception of error is influenced by the experience and imagination of observers. • Absolutely unbiased research as demanded by The International Committee of Medical Journal Editors is unattainable. • Absolute validity for evidence is an illusion.

To view Article

Introduction

Despite empirical research being based on direct observations, analytical measurements and logical interpretations of findings there is a significant potential for error, contradiction and reversal [1]. Before constituting a doctrine that applies to the community the certainty of empirical statements should be well-supported by repeated implementation studies. There is a growing need for synthesizing significance of findings from studies of varying quality using a decision framework for "best evidence" to increase transparency in systematic reviews [2,3]. Ideally, all modeling studies should include an uncertainty assessment as it pertains to the decision problem being addressed [4]. However, knowing and perceiving are interlinked and preconceived notion alters perception. This also includes the perception of error that is influenced by the experience and imagination of different researchers, reviewers and editors [5]. In addition, our imagination and our ability to think are impaired by our linguistic inaccuracy [6]. After all, science is basically constructed on unproven axioms and scientific observations are prone to deception.

Scientific investigations must meet the criteria for reliability, reproducibility and significance of results. The International Committee of Medical Journal Editors (ICMJE) promotes recommendations for creation and distribution of accurate, clear, reproducible, and unbiased articles [7]. In order to quantify the significance and rate the imprecision and inconsistency of scientific studies publications can be classified using descending degrees of assured recommendations. The Cochrane Collaboration has adopted the Grading of Recommendations Assessment, Development and Evaluation (GRADE) approach and specified four levels of certainty for a body of evidence. Assigned grades range from high and moderate to low and very low [8]. The question arises whether evidence with low and very low certainty should still be

Called evident. In this perspective I investigated the interrelations between evidence and systematic error and analyzed the possibility of unbiased research.

The body of evidence (E)

There are numerous philosophical perspectives on the relationship between theory and empirical data [9]. Historically, E in empirical research has been obtained mainly from the application of five conventional examination methods, namely observation, induction, deduction, testing and evaluation [9]. Inductive and deductive reasoning in research is performed using logical thinking. Applying deductive reasoning in the logical thinking of classical positivism, then a statement is valid when it is meaningful and verified by experience. Contrarily, Karl Popper developed a system in which falsification of the null-hypotheses can provide significance in data analysis [10]. In his model the statistical analysis of empirical data is given priority to logical thinking. In his theory of gathering scientific knowledge Popper tried a compromise defining a state between the extremes of "true" and "false" [11]. He introduced the concept that scientific progress goes along with increasing approximation to the truth, producing findings that are not true but "truelike". This agreement defines "truelike" as a state that is "true and false" at the same time. Popper`s controversial concept is based on the hypothetical thinking of the current E being closer to the truth than the precursor E. His concept does not take in to account that the formalization of the mathematical language allows the proof of everything that is knowable. This includes unrecognizable connections and error that can be proven mathematically. Kurt Gödel was able to show that number theory can prove false statements [12].

In medical science therapeutic success does not necessarily prove the correctness of diagnosis and treatment in a specific case. Therapeutic success determined by deduction promotes an impression, but does not necessarily prove it. Even an expert consensus does little to change this. This is also reflected in the following three conditions.

Condition 1: *A priori* diagnosis and treatment of a particular disease as approved in consensus expectations of experts is associated with a successful outcome (therapeutic diagnosis) in a particular case. Scientifically, this conformity proves neither the appropriateness of the diagnosis nor the efficacy of the therapy in this case.

Condition 2: *A priori* nonconventional diagnosis and treatment of a particular disease as approved in consensus expectations of experts is associated with a successful outcome in a particular case. This discrepancy proves neither the appropriateness of the diagnosis nor the efficacy of the therapy in this case.

Condition 3: An unsuccessful outcome in a specific case is confronted with an *a posteriori* corrected diagnosis and treatment by a reviewer based on approved consensus expectations of experts. This retrospective evaluation based on expert consensus does not automatically imply that the corrected diagnosis and treatment would have produced a successful outcome.

On the one hand, *Good Scientific Practice* was determined to provide the basis for the trustworthiness of scientists and their results according to professional standards, legal provisions and ethical principles. On the other hand, *Evidence-Based Medicine* advocates the conscientious and explicit use of the current "best evidence" from clinically relevant research in making decisions about the care of individual patients [13]. However, error, contradiction and reversal in empirical science can never be completely ruled out [1]. A journal's reputation as expressed in a certain amount of granted impact points might seduce some readers to lower their guard when it comes to trusting scientific paradigms. Readers should be aware of the potentially cumulative errors in systematic reviews and exercise caution when interpreting conclusions. Even Level 1 validness of E is not always the best choice or appropriate for the research question [14]. It appears that overreliance on *Evidence-Based Science* is not justified.

Error in empirical validity (V)

The current approach to *Good Scientific Practice* presumes drafting a null-hypothesis that is testable, refutable and falsifiable [9,10]. Induction is used to formulate a null-hypothesis that is based on specific observations and on existing theories, while deduction is used for testing the hypothesis. After the analysis of data and evaluation of study findings the initial null-hypothesis will then be supported or rejected [9]. The probability (P) that a null-hypothesis can be rejected is indicated by the P value and it indicates whether observed differences between groups are not due

To chance. However, V of E is limited. A falsified null hypothesis does not automatically prove an alternate hypothesis [15].

Critical rationalism encourages problem solving using trial and error. The probability of an event is a non-negative real number between the extremes of 0 and 1. As error can never be completely ruled out, P as an element of real numbers [P∈R], no matter how low the value, can never equal absolute zero. Accordingly, the P value for error is also a non-negative real number between 0 and 1 [P(e) ∈ R]. The number is always finite and always higher than zero [P(e) ∈ > 0] within the definition set of P [D={P∈R|P>0}]. The assumption that error will occur in the entire sample space (Ω) at least once is 1 [P(Ω) = 1].

In statistical analysis significance of findings is assumed when P is below a predefined significance level [α < 0.05] so that the null hypothesis can be rejected. In the binominal distribution (Bernoulli trial) the certainty of P also determines the certainty of implausibility (q) that follows the equation: q = 1 – p. When considering only random error in probability distribution, I expect the product of E and q to represented V with its presumed certainty [V = E (1 – p)]. However, this approach does not take account for diminished V from errors resulting from the choice of unsuitable methods (methodological errors). Although error is an immanent component of E not all types of error matter in a particular problem. In general, methodological errors are more detrimental than random errors.

In science the complete number of causes and their interrelations is rarely known to scientists. "Random occurrence" cannot be safely allocated unless all underlying causes are known. Whether the likelihood of error depends on the current perceptibility and conceivability of an individual observer remains a matter of debate. By all means, the observer significantly manipulates the probability of detection of identifiable errors. Similar to E, errors regarding E are subject to multiple outside influences derived from the momentary climate of time, culture, priority, ideology, methodology and linguistic ascertainment. A previously non-identifiable error may become identifiable after circumstances have changed. This is because perceptibility changes with every new experience, thus creating the potential for further new experiences. Changing perceptibility over time is an important principle of ongoing scientific research. According to Descartes, already published in 1642, doubt is the origin of all wisdom [16]. Ultimately, the impact of error depends on how a problem is seen. From this it follows that the impact of error is not absolute in nature but relative and that it strongly depends on the point of view of an individual observer. It is a hypothetical question whether all errors are potentially identifiable, as the sum of all possible errors is unlimited and may be regarded as infinite. Occasionally, identified errors are replaced by errors that have not yet been identified. The geocentric theory that prevailed until the Middle Ages saw the earth as the center of the universe. It was not until the early sixteenth century that the heliocentric theory with the sun at the center of the universe was accepted, despite massive resistance from the clergy in Europe. From today's perspective, our solar system is probably only located on the edge of a universe that we do not even know in its entire extent. We cannot tell about future corrections, which most likely will not be lasting truths either. There is no guarantee that a correction will permanently remove a recognized error.

Risk of bias (B)

Considering the variety of approaches, points of view and constructions regarding comprehensibility and perceptibility in E, there are many factors that may contribute to the risk of systematic error expressed as B. Regarding B in scientific research we can distinguish between "identified B", "identifiable but not identified B", and "non-identifiable B" (Figure 1).
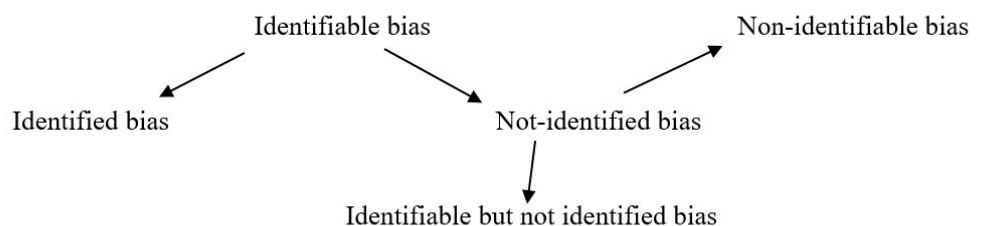

Figure 1. Different status of identification and identifiability of systematic error (bias)

B limits the significance of results. The list of cognitive and behavioral confounders that determine B is endless. Presumably, the possibility of B particularly from selection and reporting is higher in observational studies than in randomized controlled trials [17]. The Cochrane Collaboration responded with the provision of a *risk of bias tool* to rate the risk of identifiable B arising from the randomization process, deviations from intended interventions, missing outcome data, measurement of the outcome, and selection of the reported results [18]. Obviously, there are many more domains to be considered e.g. study conception, design, protocol and conduct, the study background, the financial support, the professional expertise and dependency of researchers, among others. B is ubiquitous and permanent. Joannidis defined B as the combination of various factors arising from design, data, analysis, and presentation [19]. In particular, he established that small study size, small effect size, selected test relationship, flexibility in design and definitions, and financial and scientific interests correlate with a limited truth of research findings. The risk of B in scientific studies is extremely high [19]. Apart from miscalculations, there are countless influences from B of all kinds, including association B, hindsight B, self-serving B, outcome B, availability B, authority B, overconfidence B (Dunning-Kruger effect), zero-risk B, omission B, confirmation B, selection B and self-selection B, just to mention a few of them [20]. B affects everyone involved in research, data analysis and interpretation and publication including reviewers and editors.

The probability of B particularly from selection and reporting is considered higher in observational studies than in randomized controlled trials [17]. The Cochrane Collaboration responded with the provision of a *risk of bias tool* to rate the risk of identifiable B arising from the randomization process, deviations from intended interventions, missing outcome data, measurement of the outcome, and selection of the reported result [18]. As soon as interference from B is recognized, observed B can be determined, evaluated and accounted for limitations. Occasionally, a newly observed B is replaced with another, currently not observed B. The probability of true and false applies to both, E and B.

Table 1 displays a cross tabulation of conditional probabilities based on the dependent probabilities of E and B. A probability of high (h) certainty corresponds with a low probability of error and *vice versa* a probability of low (l) certainty corresponds with a high probability of error (Tab. 1). The summation of variables in vertical direction reveals total probabilities of E and of B (P(E) = h∩E + l∩E; P(B) = h∩B + l∩B). Summation of variables in horizontal direction reveals total probabilities of h and l (P(t) = h∩E + h∩B; P(f) = l∩E + l∩B). The summation of variables in oblique direction reveals total probabilities of hE and hB (P(hE) = h∩E + l∩B; P(hB) = h∩B + l∩E).

E and B are interrelated and have to be seen in a two-way context as both go under the subjective perception of individual observers. V is composed of provable E and the currently not identified B. The sum of P(E) and P(B) makes a total of 1 [P(E) + P(B) = 1]. The relation between E and B is complementary (1).

$$E = 1 - B \qquad\qquad (1)$$

The number of all potential confounders in systematic error (B^n) is infinite (∞) and it cannot be estimated to its full extent. In this equation the exponent n would be a negative logarithmic value [$n = -\log_B (E)$]. Considering that B^n varies between individual observers, I expect the sum of all confounders in systematic error ($\Sigma\,B^n$) to lie between the threshold values n=1 and ∞ (2).

$$\sum_{n=1}^{\infty} B^n = \frac{1}{E}$$

$$(2)$$

There is an inverse relationship between B and E in which increase of B leads to decrease of E and *vice versa*. The findings of this analysis support the omnipresence of B and confirm Popper's

Chimera of a "truelike" evidence. From this point of view, absolutely unbiased research as demanded by the ICMJE is unattainable. Corresponding to error, B is an immanent components of E.

Table 1. Cross tabulation (contingency table) of actual conditions from the total probabilities of evidence P(E) and bias P(B) and of predicted conditions from the total probabilities of high accuracy P(h) and of low accuracy P(I).

		Scientific status				
		E		B		
	high	$P(h \cap E)$	$+$	$P(h \cap B)$	$=$	$P(h)$
Supposed accuracy		$+$		…..+		$+$
	low	$P(l \cap E)$	$+$	$P(l \cap B)$	$=$	$P(l)$
		$P(E)$	$+$	$P(B)$	$=$	1

The summation of variables in vertical direction reveals total probabilities of E and of B:

$$P(E) = h \cap E + l \cap E$$

$$P(B) = h \cap B + l \cap B$$

$$P(h) + P(l) = 1$$

Summation of variables in horizontal direction reveals total probabilities of h and l:

$$P(h) = h \cap E + h \cap B$$

$$P(l) = l \cap E + l \cap B$$

$$P(E) + P(B) = 1$$

Summation of variables in oblique direction reveals total probabilities of high E and high B:

$$P(hE) = h \cap E + l \cap B$$

$$P(hB) = h \cap B + l \cap E$$

$$h \cap E + l \cap B = 1$$

Conclusion

The risk for systematic error in scientific studies is extremely high. There is no absolute validity for evidence. Absolute evidence is an illusion. An illusion that we have to face. Consequently, evidence is valid as long as a finding is useful and has not been refuted by more recent studies. Validity should be supported by repeated implementations in follow-up studies rather than by a single falsification of a hypothesis. The level of evidence of a specific finding should be estimated for the accuracy of observations, measurements and calculations, and conclusions in a study. The certainty could be categorized into obvious (observational studies), proven (interventional studies) and evident (meta-analyses and systematic reviews).

Supporting information

None

Ethical Considerations

None

Acknowledgments

None

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Author contribution statement

All authors contributed equally and attest they meet the ICMJE criteria for authorship and gave final approval for submission.

Data availability statement

Data included in article/supp. material/referenced in article.

Additional information

No additional information is available for this paper.

Declaration of competing interest

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

References

1. Coccheri S. Error, contradiction and reversal in science and medicine. Europ J Int Med. *2017; 41:28-29. DOI: 10.1016/j.ejim.2017.03.026 [Crossref][PubMed][Google Scholar].*

2. Treadwell J R, Singh S, Talati R, McPheeters M L, Reston J T. A framework for best evidence approaches can improve the transparency of systematic reviews. J Clin Epidem. *2012; 65(11), 1159-1162. DOI: 10.1016/j.jclinepi.2012.06.001 [Crossref][PubMed][Google Scholar].*

3. Sarri G, Patorno E, Yuan H, et al. Framework for the synthesis of non-randomised studies and randomised controlled trials: a guidance on conducting a systematic review and meta-analysis for healthcare decision making. BMJ. *2022; 27(2), 109-119. DOI: 10.1136/bmjebm-2020-111493 [Crossref][PubMed][Google Scholar].*

4. Briggs A H, Weinstein M C, Fenwick E A, Karnon J, Sculpher M J, Paltiel A D ISPOR-SMDM Modeling Good Research Practices Task Force. Model parameter estimation and uncertainty: a report of the ISPOR-SMDM Modeling Good Research Practices Task Force--6. Value Health. 2012; 15(6), 835-842. *DOI: 10.1016/j.jval.2012.04.014 [Crossref][PubMed][Google Scholar].*

5. Lederer W. The Importance of "Nothingness" in Empirical Science – a Hypothesis. Journal of Philosophy and Ethics. *2023; 5(2):4-6. Available from: https://sryahwapublications.com/journals/journal-of-philosophy-and-ethics/volume-5/issue-2. Accessed May 24, 2024 [Article][Crossref][PubMed][Google Scholar].*

6. Wittgenstein L. Philosophical Investigations [Philosophische Untersuchungen] ed. : J. *Schulte, Frankfurt am Main, Verlag Suhrkamp, 1960. [Crossref][PubMed][Google Scholar].*

7. Recommendations for the Conduct, Reporting, Editing, and Publication of Scholarly Work in Medical Journals. Updated January 2024. Available from: https://www.icmje.org/icmje-recommendations.pdf. Accessed May 24, 2024 [Crossref][PubMed][Google Scholar].

8. Guyatt G H, Thorlund K, Oxman A D, et al. GRADE guidelines: 13. Preparing summary of findings tables and evidence profiles-continuous outcomes. *J Clin Epidemiol. 2013; 66(2), 173-183. DOI: 10.1016/j.jclinepi.2012.08.001 [Crossref][PubMed][Google Scholar].*

9. Bendassolli P F. Theory Building in Qualitative Research: Reconsidering the Problem of Induction. FQS 2013, 14(1). *[Crossref][PubMed][Google Scholar].*

10. Popper K R. The Open Society and Its Enemies: New One-Volume Edition. NED-New edition. *Princeton University Press, 1994. DOI: 10.2307/j.ctt24hqxs [Crossref][PubMed][Google Scholar].*

11. Popper K R. The growth of scientific knowledge. In: Conjectures and refutations. *London: Routledge & K. Paul, 1963. [Crossref][PubMed][Google Scholar].*

12. Gödel K. On formally undecidable theorems of the Principia Mathematica and related systems I. [Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme I]. *Monatshefte für Mathematik und Physik 1931: 38:173 198. DOI:10.1007/BF01700692 [Crossref][PubMed][Google Scholar].*

13. Sackett D, Richardson WS, Rosenberg W, Haynes RB. Evidence-Based Medicine: How to Practice and Teach EBM. London: Churchill Livingstone, 1997. *[Crossref][PubMed][Google Scholar].*

14. Ferreira P H, Ferreira M L, Maher C G, Refshauge K, Herbert R D, Latimer J. Effect of applying different "levels of evidence" criteria on conclusions of Cochrane reviews of interventions for low back pain. J Clin Epidemiol. *2002; 55(11), 1126-1129. DOI: 10.1016/s0895-4356(02)00498-5 [Crossref][PubMed][Google Scholar].*

15. Wasserstein R L, Lazar A N. The ASA Statement on p-Values: Context, Process, and Purpose. The American Statistician 2026; 70, 129-133. *DOI:10.1080/00031305.2016.1154108 [Crossref][PubMed][Google Scholar].*

16. Descartes R. Meditationes de prima philosophia in qua dei existenti et animae immortalitas demonstrater / Meditations on first philosophy (Meditationen über die Erste Philosophie). Ed. *C. Wohlers, Felix Meiner Verlag Hamburg, 2008. [Crossref][PubMed][Google Scholar].*

17. Smith G C, Pell J P. Parachute use to prevent death and major trauma related to gravitational challenge: systematic review of randomised controlled trials. BMJ. *2003; 327(7429), 1459-1461. DOI: 10.1136/bmj.327.7429.1459 [Crossref][PubMed][Google Scholar].*

18. Higgins J P, Altman D G, Gøtzsche P C. et al. Cochrane Bias Methods Group; Cochrane Statistical Methods Group. *The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. BMJ. 2011; 343:d5928. DOI: 10.1136/bmj.d5928 [Crossref][PubMed][Google Scholar].*

19. Joannidis J P A. Why most published research findings are false. PLoS Med. *2005; 2:e124. DOI:10.1371/journal.pmed.0020124 pmid:16060722 [Crossref][PubMed][Google Scholar].*

20. Wenski G. Supplement: In: Small reference book on cognitive error [Anhang. In: Das kleine Handbuch kognitiver Irrtümer]. *Springer, Berlin, Heidelberg, 2022. DOI: 10.1007/978-3-662-64776-9 [Crossref][PubMed][Google Scholar].*

Disclaimer / Publisher's Note