RESEARCH · LITERATURE AND EVIDENCE

The Hierarchy of Evidence and Its Critics

What the hierarchy of evidence is, why randomised trials and systematic reviews sit at the top, the main criticisms of ranking evidence by study design, and how frameworks such as GRADE and evidence typologies respond.

ModQu research article · 1,304 words · 13 references (APA 7th edition)

Anyone who reads research soon meets the idea that some kinds of evidence are stronger than others. The **hierarchy of evidence** is a ranking of study designs by how well they protect against bias when answering questions about cause and effect. It is usually drawn as a pyramid, with systematic reviews and randomised controlled trials at the top and expert opinion at the bottom. The hierarchy has been enormously influential, especially in medicine and public policy, but it has also attracted sustained criticism. This article sets out what the hierarchy claims, where it comes from, what its critics say and how more recent approaches try to keep its strengths while avoiding its weaknesses.

What the hierarchy claims

The hierarchy grew out of the evidence-based medicine movement of the 1990s. Sackett et al. (1996) defined evidence-based medicine as the conscientious, explicit and judicious use of current best evidence in making decisions about the care of individual patients. Best evidence, for questions about whether a treatment works, meant evidence from studies designed to rule out alternative explanations. Rankings of study designs made this practical: a busy clinician could look first for the designs most likely to give an unbiased answer.

A typical version of the hierarchy is shown below.

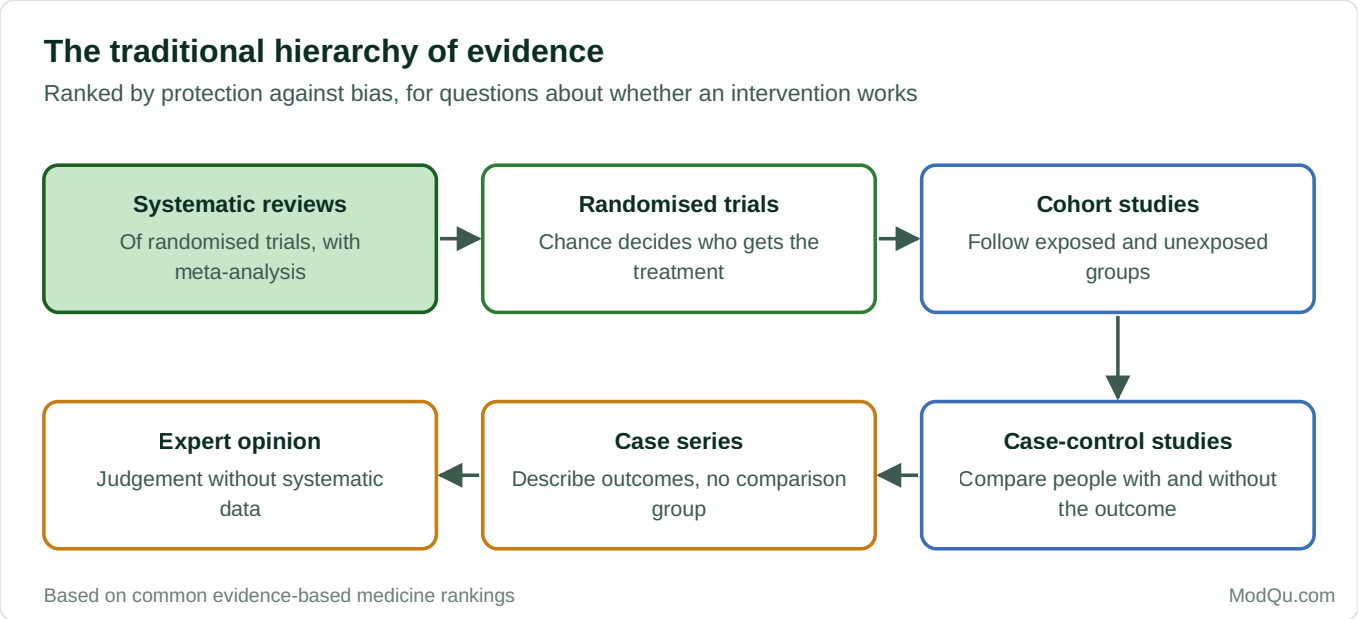


Figure 1. A typical version of the traditional hierarchy of evidence for questions about whether an intervention works. Based on common evidence-based medicine rankings; diagram by ModQu.

The logic is straightforward. In a randomised trial, chance decides who receives the treatment, so the groups being compared should differ only by chance in everything other than the treatment, including factors nobody thought to measure. In a cohort or case-control study, people choose or are assigned to exposures in ways that may be linked to their outcomes, so differences between groups may reflect those links rather than the exposure itself. Case reports and expert opinion offer no comparison group at all. A systematic review, which brings together all the relevant trials using explicit methods, sits above any single trial because it reduces the influence of chance and of selective attention to particular studies (Higgins et al., 2019).

Why the hierarchy matters

The hierarchy has done real good. Ioannidis (2005) showed that many highly cited findings from observational studies were later contradicted or found to be exaggerated when tested in randomised trials. Ranking designs gave practitioners a defensible reason to trust some claims less than others, and to resist conclusions based on authority, tradition or a few memorable cases. It also gave research funders a clear signal about what kinds of studies were needed to settle important questions.

The main criticisms

Critics do not usually argue that study design is irrelevant. Their objection is that a single ranking, applied mechanically, can mislead. Five criticisms recur.

It answers only one kind of question. The hierarchy ranks designs for questions about whether an intervention causes an outcome. Many important questions are of a different kind: how common a problem is, what people's experiences are, why a programme is accepted or resisted, or how a process actually works in practice. Petticrew and Roberts (2003)

argued that the right design depends on the question being asked, and that for many questions a survey, a qualitative study or a cohort is the best available design, not a weaker substitute for a trial. They described the alternative as an evidence **typology**: a matching of questions to designs, rather than a single ladder.

Design is not the same as quality. A poorly conducted trial, with high drop-out, unblinded outcome assessment or selective reporting, may be less trustworthy than a careful observational study. Ranking by design label alone ignores how well each study was done.

Observational studies are not always wrong. Benson and Hartz (2000) and Concato et al. (2000) compared the results of well-designed observational studies with those of randomised trials on the same questions and found that, in many cases, the estimates were similar. They concluded that the assumption that observational studies systematically overestimate effects was not supported for the topics they examined, although both groups stressed that the result depended on good design.

Some questions cannot be answered by trials. Smith and Pell (2003) made this point with deliberate humour, reporting that no randomised trials had tested whether parachutes prevent death when jumping from aircraft. Where effects are very large, where randomisation would be unethical, or where outcomes are rare or long-delayed, other designs must carry the weight.

Internal validity is not external validity. A randomised trial shows that a treatment worked, on average, in the people and setting studied. Rothwell (2005) showed that trials often exclude many of the patients who later receive the treatment and that reports seldom give enough detail for users to judge whether results apply to them. Deaton and Cartwright (2018) argued that the language of a "gold standard" encourages people to forget this gap. Cartwright and Hardie (2012) put the point practically: to predict whether a policy will work in a new place, decision-makers need to know why it worked where it was tested and whether the supporting conditions are present in the new setting. Pawson and Tilley (1997) built a whole approach to evaluation on asking what works, for whom, in what circumstances and why, rather than simply whether something works.

Responses: from ranking to rating

These criticisms have changed practice. The most widely used response is the GRADE approach (Guyatt et al., 2008). GRADE still starts from study design: evidence from randomised trials begins as high certainty and evidence from observational studies begins as low certainty. But the starting point can then move. Trial evidence is rated down for risk of bias, inconsistency between studies, indirectness, imprecision and suspected publication bias. Observational evidence can be rated up when effects are large, when there is a dose-response gradient, or when plausible biases would have worked against the observed effect. The result is a judgement about the certainty of a body of evidence for a specific question, not a fixed ranking of designs.

The second response is the typology approach, which matches designs to questions, as summarised in the table below.

Horses for courses

The best design depends on the question being asked

Question	Example	Best-suited designs
Effectiveness	Does the programme change behaviour?	Randomised trials, quasi-experiments
Process	How is the programme delivered?	Qualitative studies, case studies
Acceptability	Do people want it?	Qualitative studies, surveys
Prevalence	How common is the problem?	Cross-sectional surveys
Cost	Is it worth the money?	Economic evaluations
Experience	What is it like for users?	Interviews, ethnography

Adapted from Petticrew & Roberts (2003) ModQu.com

Figure 2. Matching research questions to study designs. Adapted from Petticrew and Roberts (2003); table by ModQu.

In the social sciences, where interventions are often complex and outcomes varied, Petticrew and Roberts (2006) argued that systematic reviews should include evidence from many designs, including qualitative research, and should be explicit about how each kind of evidence contributes to the answer.

Using the hierarchy well

For anyone reading or commissioning research, including business research, a few practical rules follow from this debate.

- **Start with the question.** Decide first whether the question is about cause and effect, prevalence, experience, process or something else, and then ask which designs are best suited to it.
- **Judge the study, not only its label.** A design that is strong in principle can be weak in practice. Look at sampling, measurement, missing data and reporting.
- **Weigh the body of evidence.** One study rarely settles anything. Consistent findings from several designs with different weaknesses are more convincing than a single study at the top of the pyramid.
- **Ask whether results will travel.** Consider whether the people, setting and conditions of the original study resemble the situation in which the results will be used.

Conclusion

The hierarchy of evidence remains a useful reminder that some designs protect better against bias than others when the question is whether something works. Its critics have shown that it is not a universal ranking of all evidence for all purposes. The more defensible position, reflected in GRADE and in evidence typologies, is to match designs to questions, judge each study on how well it was done, and rate the certainty of the whole body of evidence rather than trusting a label.

References

Benson, K., & Hartz, A. J. (2000). A comparison of observational studies and randomized, controlled trials. *New England Journal of Medicine*, *342*(25), 1878–1886. https://doi.org/10.1056/NEJM200006223422506

Cartwright, N., & Hardie, J. (2012). *Evidence-based policy: A practical guide to doing it better*. Oxford University Press.

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*(25), 1887–1892. https://doi.org/10.1056/NEJM200006223422507

Deaton, A., & Cartwright, N. (2018). Understanding and misunderstanding randomized controlled trials. *Social Science & Medicine*, *210*, 2–21. https://doi.org/10.1016/j.socscimed.2017.12.005

Guyatt, G. H., Oxman, A. D., Vist, G. E., Kunz, R., Falck-Ytter, Y., Alonso-Coello, P., & Schünemann, H. J. (2008). GRADE: An emerging consensus on rating quality of evidence and strength of recommendations. *BMJ*, *336*(7650), 924–926. https://doi.org/10.1136/bmj.39489.470347.AD

Higgins, J. P. T., Thomas, J., Chandler, J., Cumpston, M., Li, T., Page, M. J., & Welch, V. A. (Eds.). (2019). *Cochrane handbook for systematic reviews of interventions* (2nd ed.). Wiley. https://doi.org/10.1002/9781119536604

Ioannidis, J. P. A. (2005). Contradicted and initially stronger effects in highly cited clinical research. *JAMA*, *294*(2), 218–228. https://doi.org/10.1001/jama.294.2.218

Pawson, R., & Tilley, N. (1997). *Realistic evaluation*. SAGE.

Petticrew, M., & Roberts, H. (2003). Evidence, hierarchies, and typologies: Horses for courses. *Journal of Epidemiology and Community Health*, *57*(7), 527–529. https://doi.org/10.1136/jech.57.7.527

Petticrew, M., & Roberts, H. (2006). *Systematic reviews in the social sciences: A practical guide*. Blackwell. https://doi.org/10.1002/9780470754887

Rothwell (2005) showed that randomised trials in medicine often exclude many of the patients who will later receive the treatment, such as older people or those with several conditions, and that reports seldom give enough detail for clinicians to judge whether the results apply to their own patients.

Sackett, D. L., Rosenberg, W. M. C., Gray, J. A. M., Haynes, R. B., & Richardson, W. S. (1996). Evidence based medicine: What it is and what it isn't. *BMJ, 312*(7023), 71–72. https://doi.org/10.1136/bmj.312.7023.71

Smith, G. C. S., & Pell, J. P. (2003). Parachute use to prevent death and major trauma related to gravitational challenge: Systematic review of randomised controlled trials. *BMJ, 327*(7429), 1459–1461. https://doi.org/10.1136/bmj.327.7429.1459