Legislative Council Staff
Nonpartisan Services for Colorado's Legislature

Memorandum

January 26, 2026

TO: Interested Persons

FROM: Samantha Lattof, PhD MSc, Science and Technology Policy Program Fellow

SUBJECT: Understanding the Hierarchy of Evidence: A Framework for Health Research

Table of Contents

Overview

This memo details the hierarchy of evidence framework, a tool used in health research to rank studies based on their ability to minimize bias and establish trustworthy conclusions. It reviews the strengths and weaknesses of different types of health research studies. Finally, it offers strategic guidance for applying the hierarchy to evidence-based health policymaking, emphasizing the need to balance study design with real-world feasibility and approaches for assessing recommendation strength.

Background

Accurately answering a question about a health intervention can involve sorting through thousands of studies. Studies may use different methods and different types of data, making it challenging to compare findings. To help inform clinical decision making and ensure that health interventions are safe and effective, scientists have developed hierarchies and systems to distinguish stronger forms of research from weaker forms of research.

The first recognizable evidence hierarchy, published in 1979, established three levels of evidence. Evidence from randomized controlled trials appeared at the top, followed by evidence from well-designed cohort or case-control studies, and then opinions of respected clinical authorities or reports from expert committees. Critical appraisal strategies, problem-based learning methodology, and the Evidence-Based Medicine movement led to refinements in the framework. A proliferation of versions followed, remaining true to the hierarchy's core structure but adopting a pyramidal form and reflecting author preferences (e.g., the inclusion of animal studies and in vitro studies).

Researchers use the evidence hierarchy framework to help:

- identify what information is more or less trustworthy;
- make informed decisions;
- avoid being misled; and
- produce analyses that people can act on with greater confidence.

While the evidence hierarchy is a useful tool, it has limitations. No "perfect" study type exists. Researchers must choose a study design that most closely fits their specific research question (e.g., "Does this drug work?" versus "Is this rare chemical toxic?"). A randomized controlled trial, for example, may present with a high level of evidence on the pyramid (e.g., "There is high evidence the drug works."). However, that placement in the hierarchy does not mean that type of study will always produce a strong recommendation (e.g., "While the drug works, it causes severe side effects, leading to a 'weak' recommendation."). Ethical considerations may also limit the ability to conduct a particular type of research in a particular population.

Despite the pyramid's distinct tiers, the lines between these tiers should be considered wavy rather than rigid. A large, well-designed, and properly conducted cohort study may provide more reliable evidence than a single, small, or poorly executed randomized controlled trial. For instance, observational studies established the link between smoking and lung cancer rather than randomized controlled trials.

While the evidence hierarchy provides clarity around how a study was built, evaluation systems like the Grading of Recommendations Assessment, Development, and Evaluation (GRADE) approach provide information about study quality, how much people can trust studies' results, and whether or not they should act on those results. Despite these crucial differences, the evidence hierarchy remains an essential starting point for scientific literacy.

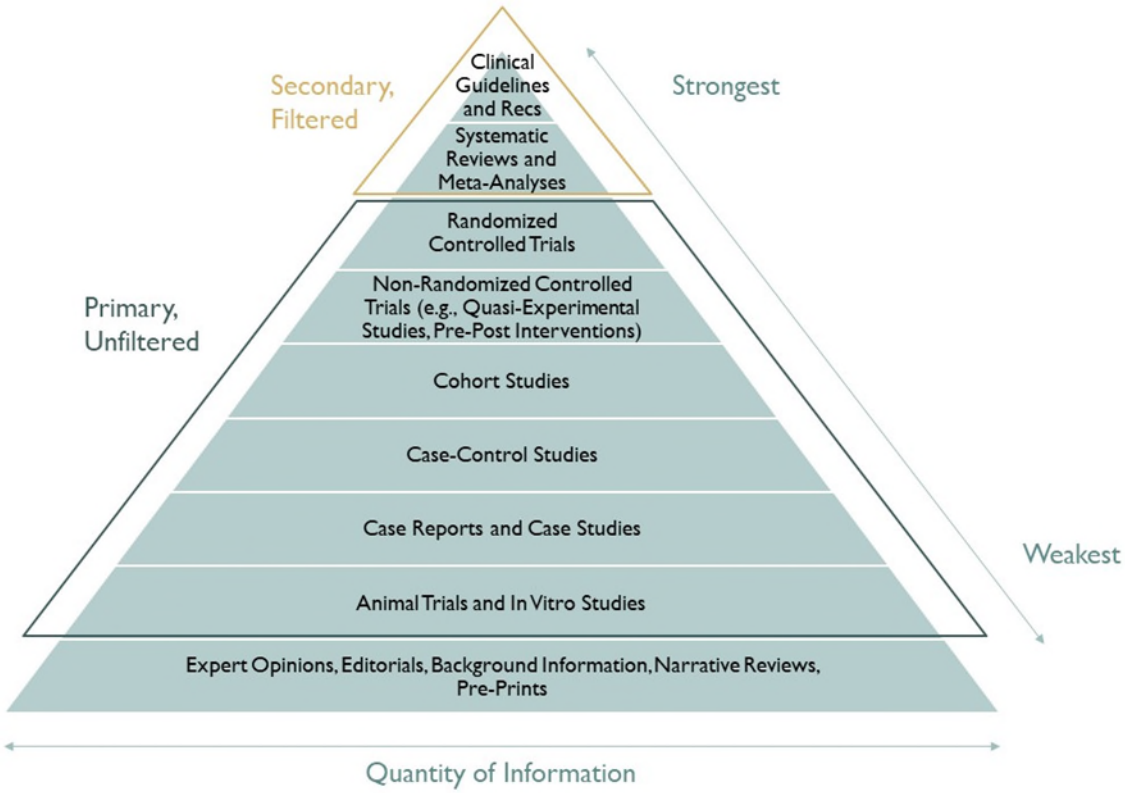
The Hierarchy of Evidence

The hierarchy of evidence is depicted as a pyramid, with strongest forms of evidence at the top and weakest forms of evidence at the bottom (Image 1). The top of the pyramid prioritizes high-quality research with more robust study designs and minimized potential for bias, as these types of research offer stronger conclusions. Information in the top two levels relies on secondary data[1] sources that researchers have filtered using techniques to weed out low-quality evidence.

Towards the middle of the pyramid are studies that produce unfiltered, primary data.[2] Most studies in the middle are observational studies (e.g., cohort studies, case-control studies, cross-sectional studies, case series), meaning that researchers are observing and recording events without imposing any intervention. While the sources of evidence towards the bottom of the pyramid tend to be most widely available, they are also the weakest for generating strong conclusions.

[1] Secondary data refers to data that were already collected by other researchers.
[2] Primary data refers to original, first-hand data that researchers collected directly.

Figure 1
The Hierarchy of Evidence for Health Research



Source: Legislative Council Staff.

The following sub-sections briefly review these different types of evidence in order of strongest to weakest forms.

Clinical Practice Guidelines and Recommendations

Clinical practice guidelines and recommendations provide consensus recommendations developed by panels of experts who are usually informed by systematic reviews of the best available evidence and a robust evaluation of the evidence.[3] As these guidelines and recommendations are often the most authoritative statements on what works in health and medicine, their findings are of particular importance to policymaking.

[3] The World Health Organization's 2025 Guideline for the prevention, diagnosis and treatment of infertility is an example of an evidence-informed guideline that was developed using a rigorous methodological process. Section 2.6 in the guideline clearly lists the methods used to review evidence, to make recommendations, and to manage potential conflicts of interest by the experts involved in the process.

Strengths

Clinical practice guidelines and recommendations are considered strong forms of evidence because they:

- synthesize large bodies of primary research;
- are developed using rigorous, transparent methodologies and evaluation processes;
- minimize bias by requiring experts involved in the development to disclose conflicts of interest;
- take patients' real-world experiences, preferences, and actions into account; and
- prevent "cherry-picking" of data or a single study that supports a specific political agenda.

Limitations

- If the guidelines or recommendations do not address potential conflicts of interest by the experts involved, proceed with caution. The resulting product may reflect the biases of these individuals and the entities that fund them.
- If the guidelines or recommendations do not explain their methodology, proceed with caution. Cherry-picking of evidence may result in recommendations with low confidence.

Systematic Reviews and Meta-Analyses

Systematic reviews methodically collect *all* good studies on a question, whereas meta-analyses combine data statistically to show overall effects. Researchers often combine these methods into a combined product, reporting on a systematic review and meta-analysis in a single paper.[4]

To ensure systematic reviews and meta-analyses are robust, the authors' methods should provide transparency about how the included studies were chosen and the tool(s) the authors used to quality assess included studies (e.g., AMSTAR 2, Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA)).

Strengths

Systematic reviews and meta-analyses are considered strong forms of evidence because they:

- synthesize large bodies of primary research;

[4] Tamblyn and colleagues' 2022 article "Vitamin D and miscarriage: a systematic review and meta-analysis," published in Fertility and Sterility, is an example of a systematic review and meta-analysis conducted using well-established processes. The authors registered their research with the International Prospective Register of Systematic Reviews, reported their work following the PRISMA guidelines, assessed quality using tools appropriate for the study types, and disclosed their interests and funding sources. They also employed the GRADE approach to evaluate the confidence in evidence for each intervention.

- provide more precise estimates of effects than any single study due to the pooling of data from multiple studies;
- reduce bias by looking at the full body of high-quality evidence;
- help clarify why different studies on the same topic may have come to different conclusions; and
- use pre-defined, transparent methodologies to find and assess all available evidence, reducing the risk of "cherry-picking" evidence that supports a specific political agenda.

Limitations

- The quality of these studies depends on the researchers' techniques, the relevance of the research question, and the authors' methodological rigor (e.g., their inclusion criteria, the depth of their searches, their risk of bias evaluations).
- Publication bias from not publishing research with negative or boring results means the systematic review might falsely suggest an intervention is more effective.
- These studies can become outdated as new data emerges, since they are time-consuming to conduct.

Randomized Controlled Trials

Randomized controlled trials[5] randomly assign participants to receive an intervention or not (e.g., drug A vs. a placebo, intervention B vs. no intervention), helping isolate cause and effect. This study design is widely considered to be the "gold standard" for causality, testing whether or not an intervention works.

Strengths

Strengths of randomized controlled trials include that they:

- help minimize bias and confounding[6] by design, particularly studies using a "double-blind" method (neither patient nor doctor knows who received the treatment); and
- ensure any difference in outcomes is likely due to the treatment, not hidden differences between participants, due to randomization.

---

[5] Jastreboff and colleagues' 2022 article, "Tirzepatide Once Weekly for the Treatment of Obesity" in the New England Journal of Medicine, is an example of a randomized controlled trial. By using a double-blind approach, the researchers could prove that the significant weight loss they observed was due to tirzepatide rather than to people's expectations or bias.

[6] Confounding occurs when a variable that is not the outcome or the exposure affects the "real" effect of the outcome on the exposure.

Limitations

Limitations of randomized controlled trials include that they:

- can be expensive and slow to conduct;
- may not be ethical or feasible to conduct; and
- may not reflect real-world conditions that could be important for policymaking (e.g., often exclude participants with multiple health issues).

Non-Randomized and Quasi-Experimental Studies

Non-randomized and quasi-experimental studies[7] evaluate interventions without full randomization, often using real-world data (e.g., natural experiments where one state adopts a law and another does not, interrupted time series, pre-post designs). A quasi-experimental approach is the standard for evaluating policy impact since researchers cannot randomly assign a state or country to pass a law.

Strengths

Strengths of non-randomized or quasi-experimental studies include that they:

- are often the best type of evidence available for real-world policy evaluations, like evaluating the impact of a new city-wide law or hospital-wide protocol; and
- have increased generalizability since they occur in real-world settings, where the results are likely to reflect what actually happens in practice (as opposed to the artificial control offered by a randomized controlled trial).

Limitations

Limitations of non-randomized or quasi-experimental studies include that they:

- may vary in strength from moderate to strong, depending on the study design and assumptions;
- may have skewed results from selection bias (e.g., health-conscious people choose to join the new wellness program); and
- are more susceptible to confounding than randomized controlled trials.

---

[7] Pell and colleagues' 2008 article, "Smoke-free Legislation and Hospitalizations for Acute Coronary Syndrome" in the New England Journal of Medicine, is an example of a quasi-experimental study. Using an interrupted time-series design, the researchers measured the number of hospital admissions for acute coronary syndrome before and after Scotland's implementation of a national law prohibiting smoking in all enclosed public places.

Cohort Studies

Cohort studies[8] follow groups of people over time to see how exposures relate to outcomes. They are widely considered to be the strongest type of observational study. This prospective study design is excellent for tracking disease progress or the long-term effects of lifestyle choices over decades. Cohort studies may be prospective, following people into the future, or they may be retrospective, reviewing data from medical charts to understand the past.

Strengths

Strengths of cohort studies include that they:

- can help establish a timeline, proving that the exposure occurred before the disease developed;
- are efficient in studying multiple outcomes simultaneously through a single cohort; and
- help study real-world exposures and detect harms or long-term effects.

Limitations

Limitations of cohort studies include that they:

- cannot fully prove cause-and-effect;
- can require long follow-up lasting decades, leading to high costs of tracking people over the time period;
- may have skewed results due to hidden factors (e.g., wealthy people can afford better diets); and
- may lose participants to follow-up since the studies often take decades.

Case-Control Studies

Case-control studies start with people with a particular outcome and compare them to those who do not, looking back in time for differences. They are usually, but not exclusively, retrospective. These studies are ideal for studying new outbreaks or rare diseases because they start with people who are already sick.

[8] Klein and colleagues' 2025 article, "Two years and counting: a prospective cohort study on the scope and severity of post-COVID symptoms across diverse patient groups in the Netherlands—insights from the CORFU study" in BMJ Open, is an example of a cohort study. The authors also registered their study and published their protocol. These practices enable other scientists to replicate the study and evaluate the authors' methodology against best practices. Following a documented protocol also helps minimize risk of bias and selective reporting.

Strengths

Strengths of case-control studies include that they:

- are efficient for studying rare diseases and outbreaks; and
- are useful for identifying potential risk factors.

Limitations

Limitations of case-control studies include that they:

- cannot establish causality;
- cannot determine how common the disease is in the general population (incidence) though they can show a link;
- may have skewed results due to hidden factors (e.g., wealthy people can afford better diets); and
- may experience retrospective study design biases resulting from limitations like missing information in medical charts, challenges with recall, and unknown reasons for patient selection.

Case Reports and Case Studies

Case reports[9] describe a single patient or event, whereas case studies[10] describe a small group of patients or events. These studies are often the first way that people discover new diseases (e.g., SAR-CoV-2, AIDS) or unexpected side effects of a drug.

Strengths

Strengths of case reports and case studies include that they:

- provide early warnings of new diseases or emerging issues (e.g., unexpected drug side effects); and
- generate hypotheses for more rigorous research studies by providing initial clues.

---

[9] Holshue and colleagues' 2020 article, "First Case of 2019 Novel Coronavirus in the United States" in the New England Journal of Medicine, is an example of a case report. The authors reported the first case of 2019-nCoV infection confirmed in the United States, describing "the identification, diagnosis, clinical course, and management of the case, including the patient's initial mild symptoms at presentation with progression to pneumonia on day 9 of illness."

[10] Guan and colleagues' 2020 article, "Clinical Characteristics of Coronavirus Disease 2019 in China" in the New England Journal of Medicine, is an example of a case series. The authors reported findings from 1,099 patients with laboratory-confirmed Covid-19 in an attempt to help identify the disease's defining clinical characteristics and severity.

Limitations

Limitations of case reports and case studies include that they:

- lack a control group, making it impossible to prove the outcome was caused by the treatment rather than pure chance;
- have high bias; and
- are not generalizable.

Types of Evidence at the Bottom of the Pyramid

Animal Trials and In Vitro Studies

Animal trials and in vitro studies produce foundational biomedical research required in the development of new drugs and treatments. However, just because an intervention works in animals or in a lab, that does not mean it will produce the same results in humans.

Journalists and press releases have frequently conflated findings in mice to takeaways in humans, writing headlines like, "Exercise during pregnancy protects children from obesity, study finds: New research suggests lack of exercise in healthy women during pregnancy can predispose their children to obesity." If a reader only skimmed the headline, they would never have known those findings resulted solely from an animal trial in mice.

Expert Opinions and Editorials

Expert opinions and editorials help fill gaps by providing essential guidance when no high-level research yet exists, such as at the beginning of the SARS-CoV-2 pandemic. However, they are the most prone to personal biases, financial conflicts of interest, and antiquated thinking (e.g., "We've always done it this way.").

Expert opinions at the bottom of the pyramid and expert consensus used in the development of clinical guidelines and recommendations at the top of the pyramid have two crucial differences:

1. transparency in developing the guidelines and recommendations; and
2. the systematic nature of the methods and the research used to reach expert consensus in the guideline and recommendation development process.

Background Information and Narrative Reviews

Background information and narrative reviews can be helpful in developing reading lists, but these papers do not collect evidence in a structured, methodical way like systematic reviews. As a result, background information and narrative reviews are prone to author biases and may reflect an incomplete picture.

Preprints

Preprints are preliminary drafts of scientific research papers that have not been through the scrutiny of editorial and peer-review processes required for publication in a reputable journal. Since publishing in a journal can take months or years, posting preprints online makes findings more quickly available and can be beneficial during the emergence of new diseases (e.g., SARS-CoV-2). However, until the research has been through the editorial processes and published in a peer-reviewed journal, this source of evidence is best suited for raising awareness of timely research results so that scientists can establish priority for future research.

Tips for Applying the Hierarchy to Evidence-Based Health Policymaking

While the evidence hierarchy can help guide the translation of science into policy, its application requires a certain degree of nuance. The following tips may help applying the hierarchy of evidence to health policymaking.

- Prioritize evidence from high-level syntheses like clinical guidelines, systematic reviews, and meta-analyses. They save time by providing the big picture quickly. These types of evidence also help prevent "cherry-picking" of a single study selected to support a specific political agenda.
- When evidence conflicts, check the study types. For example, findings from a randomized controlled trial generally outweigh those from a single observational study. A systematic review generally outweighs both.
- Context matters. While the hierarchy ranks randomized controlled trials high, this type of research is often conducted in controlled settings that may differ from how a health intervention works in the "real world." Looking at implementation science and observational data may provide policy-relevant information to complement the findings.
- When examining policy impacts, look for quasi-experimental studies. These study types are often the best real-world evidence about laws, programs, and state-level interventions.
- Identify when the evidence is too weak to support strong claims, as it will help avoid overstated or premature conclusions.
- Note gaps honestly. If only case reports or expert opinions exist, it is alright to say so. Apart from helping build credibility, it shows there is still a degree of uncertainty.

- That said, an absence of high-level evidence does not mean an intervention is ineffective. It may simply have not yet been studied that way. Rare diseases and specialized interventions often lack randomized controlled trials because it would be unethical or logistically impossible to conduct them.
- In public health emergencies, like the beginning of a pandemic, policymakers may need to act on "weak" forms of evidence (e.g., case study, expert opinion) because waiting for stronger research studies and analyses could result in harm.
- While it is important to understand how studies are built, identifying strong forms of evidence for policymaking should go beyond study design to include factors like the magnitude of the effect, the uncertainty of the results, and the feasibility of the policy (Guyatt et al. 2008). Consider evidence that has been quality assessed and rated for recommendation strength using tools like the GRADE approach.
- The hierarchy of evidence does not account for patient experiences, patient preferences, and patients' actions. Real-world evidence from patients can provide crucial insights for policy.