

4.5 Distinguishing high- from low-quality evidence

Not all evidence is high quality and reliable for making decisions. Tools exist for many (but not all) forms of evidence to help make judgments about the quality of the evidence (from a single study or a body of evidence) can be relied upon. As we describe here, these tools use scores or grades to help users understand how confident they can be in the evidence. Many journals now require authors to follow reporting standards, such as CONSORT for randomised-controlled trials and PRISMA for evidence syntheses. Most journals do not require reviewers to use specific tools to assess the quality of studies or strength of recommendations; as a result, publication in a peer-reviewed journal is not a good proxy for quality.

Issue	Response
Studies (and guidelines) vary in their quality (or trustworthiness)	- Quality-assessment (or critical-appraisal) tools have been developed for specific study designs (e.g., randomised-controlled trial), for broad categories of study designs (e.g., observational study, qualitative research, and evidence synthesis), and for guidelines – see the annex at the end of this chapter (section 4.16) for examples (RoB2, ROBINS-I, JBI checklist, AMSTAR, and AGREE II) - Tools may yield a summary judgment (e.g., low risk of bias using RoB2 or ROBINS-I), a score (although some groups inform ranges (e.g., high quality using AMSTAR), a set of scores (e.g., six domains using AGREE II), or a set of considerations that can inform a summary judgment (e.g., JBI checklist)
Bodies of evidence vary in their certainty (or the confidence you can place in them)	- Certainty-assessment tools have been developed for a body of evidence addressing the same question (e.g., effect of an intervention on a specific outcome or the meaningfulness of a clinical trial) – see section 4.16 for more examples (GRADE and GRADE CERQal) - Tools may yield a summary judgment about confidence that the true effect is similar to the estimated effect (e.g., high certainty in GRADE) or that the phenomenon of interest is well represented by a qualitative study finding (in GRADE CERQal) - A summary judgment about the certainty of an effect estimate is more helpful than a test of statistical significance demonstrating that an intervention ‘works’ or ‘doesn’t work’ (which will happen by chance one in 20 times if statistical significance is set at the 0.05 level)
Recommendations vary in their strength	- Strength-assessment tools have been developed for guideline recommendations (e.g., GRADE, in addition to ranking the certainty of a body of evidence, as described above) – see section 4.16 for an example - Tools may yield a summary judgment about whether most decision-makers would choose to proceed with an intervention (e.g., strong in GRADE) or whether most would need to carefully weigh the pros and cons of an intervention
Some sources of (or approaches used to generate) evidence can be hard to judge	- No ideal accepted tools exist to assess how much confidence can be placed in: - An expert, although examples like The Good Judgment Project do exist for forecasting (see more on expert opinion later in this chapter and, in the case of expert opinion about model parameters, in section 4.16) - Models used in generating some forms of evidence (which we address in section 4.7 when talking about climate-change models and in section 4.16) - An artificial-intelligence algorithm used in generating some types of evidence, although examples like TRIPOD are starting to emerge (3)

Distinguishing high- from low-quality evidence is particularly challenging when evidence is embedded in dashboards, models and other formats, and when conflicts of interest are at play. We return to the latter in [sections 4.12, 4.14 and 4.16](#). While not the focus of this report, distinguishing high- from low-quality ‘radar’ can also be challenging, and organisations like UNICEF have developed data-quality frameworks to assist with this (bit.ly/3DQQRrv).

Some ‘one-stop shops,’ such as Social Systems Evidence and the COVID-19 Evidence Network’s *Support Decision-making (COVID-END)* in *enfor* (described in [section 4.6](#)), use some of these tools so that decision-makers and those supporting them can focus on high-quality evidence synthesis or *nders* and that they are using the best available (if not high-quality) evidence synthesis.

The COVID-19 pandemic required decision-makers to make difficult decisions in short time frames, initially with little and often indirect evidence, and then, over time, with studies, bodies of evidence, and recommendations developed using a robust process. To support decision-making about COVID-19 based on bodies of evidence (rather than single studies), COVID-END profiled in its *enfor* of ‘best’ evidence synthesis that were *p-to-date* (based on the date of searching for evidence), were high quality (based on the AMSTAR tool), and provided an assessment of the certainty of the evidence (based on the GRADE tool).

Just as not all evidence is high quality, not all global evidence will be applicable in a given context. For example, an evidence synthesis containing studies conducted in only high-income countries may have limited applicability to some low-income countries. There may be important differences in baseline conditions, in on-the-ground realities and constraints, and in structural features of the local system (e.g., national health system or provincial/state education system). A SUPPORT tool can also help people think through the local applicability of findings from an evidence synthesis and consider how insights can still sometimes be drawn even when the findings aren’t applicable.⁽⁴⁾

Bayesian reasoning has garnered increasing attention as a way to deliberate the ‘re-drawing of mental maps’ about challenges and ways of addressing them, not by replacing all of what we know with new information, but by modifying our understanding to an appropriate degree. The degree depends on how much confidence we had in our pre-existing knowledge (the ‘prior’ probability of something being true) and how much confidence we place in the new knowledge. More confidence can be placed in the new knowledge. More confidence can be placed in the new knowledge.