4.16 Annex to section 4.5 – Examples of quality-assessment tools

Form of evidence	Examples of quality-assessment tools
	Types of evidence for which quality-assessment tools exist
Data analytics	ROBINS-I (riskofbias.info) for observational studies, such as those that examine associations between select factors (including interventions) and select outcomes, where there is a risk of bias from: • confounding (where the observed relationship between a factor and an outcome, differs from the true relationship because of one or more additional factors that are not accounted for) • selection of participants into the study • classification of intervention(s) • deviations from intended intervention • missing data • measurement of outcomes • selection of the reported result
Evaluation	Risk of Bias (RoB) 2 (riskofbias.info) for randomized-controlled trials, where the risk of confounding is less, but where there is a risk of bias from some (albeit fewer) of the same sources as above: • randomization process • deviations from the intended interventions • missing (outcome) data • measurement of outcomes • selection of the reported result
Behavioural/ implementation research	See other rows for the relevant types of studies or syntheses
Qualitative research	JBI critical appraisal checklist for qualitative research (bit.ly/31Lsib1), where very different considerations come into play, such as: • congruity between the research methodology and the research question, data-collection methods, data representation and analysis, and results interpretation, as well as between the stated philosophical perspective and the methodology • reflexiveness on the part of the researcher, such as statements locating the researcher culturally and theoretically, and addressing the researcher's influence on the research and vice versa • representation of study participants and their voices • flow of conclusions from the analysis and interpretation of the data
Evidence synthesis	See above for the relevant types of studies considered in the evidence synthesis
	A MeaSurement Tool to Assess systematic Reviews (AMSTAR; amstar.ca) for the quality of the evidence synthesis, where the risk of bias can arise from: • identification of all potentially relevant studies through a comprehensive search of both published and grey literature and without language restrictions • selection of all studies addressing the research question using explicit criteria about study designs and about participants, interventions/factors, comparisons and outcomes, and with at least two reviewers applying the criteria • quality appraisal of and data extraction from all included studies • synthesis of findings from all included studies Note that there are two versions of AMSTAR: 1) the original version that can be applied across all types of syntheses, albeit with some criteria removed from both the numerator and denominator; 2) a second version of AMSTAR that is more specifically relevant to syntheses of randomized-controlled trials
	Grading of Recommendations, Assessment, Development and Evaluations (GRADE; bit.ly/3C9pMrx) for the certainty of evidence for the outcomes of an intervention, with: • certainty rated down because of risk of bias (with evidence from randomized-controlled trials starting at high certainty and evidence from observational studies starting at low quality and then being adjusted based on RoB2 or ROBINS-I), imprecision (e.g., one or two small studies), inconsistency (e.g., two studies showing very different findings), indirectness (e.g., surrogate measures used or atypical settings studied), and publication bias (e.g., more common with observational studies because of the lack of study registries or with industry-funded studies because of the commercial incentive to publish positive studies) • certainty rated up for large magnitude of effect, dose-response gradient, and when all residual confounding would decrease the magnitude of effect
	GRADE CERQual (cerqual.org) for the certainty of evidence for the qualitative representation of a phenomenon of interest, with: • certainty rated down because of concerns about methodological limitations (because problems in the way studies were designed or reported were identified using a critical-appraisal tool like the JBI one above), relevance (because the context in which the primary studies were conducted are substantively different from the context of the synthesis question), coherence (because some of the data contradict the findings or are ambiguous), and adequacy (because the data are not sufficiently rich or only come from a small number of studies or participants)

	Technology assessment / cost-effectiveness analysis	International Network of Agencies for Health Technology Assessment (INAHTA) checklist (bit.ly/2YJVMVK) for the quality of technology assessments, with two of the 14 questions addressing the approach to synthesizing the evidence (with prompts similar to AMSTAR) and another question addressing whether the assessment was contextualized through an accompanying cost-effectiveness analysis (with local – meaning national or sub-national – costing data), and consideration of local legal, ethical and social implications
		Drummond checklist of cost-effectiveness analyses (bit.ly/3FbnB8R), and for economic evaluations more generally, with questions about study design, data collection, and the analysis and interpretation of results
		Philips checklist for cost-effectiveness analyses that include a decision-analytic modeling component (bit.ly/3FcWBGc) with questions about the structure of the model (e.g., explicit rationale, justified assumptions and appropriate time horizon), the data used (e.g., baseline probabilities from observational studies, treatment effects from randomized-controlled trials, and assessments of four types of uncertainty, namely the structure of the model, the methodological steps followed, the heterogeneity in the population studied, and the parameters used), and the consistency (internal and external) – there is also the complementary TRUST tool to assess uncertainties in decision-analytic models (bit.ly/3quFSKp)
	Guidelines	AGREE II tool (bit.ly/30qyFAb) for assessing the development, reporting and evaluation (or quality appraisal) of guidelines, which uses 23 items grouped into six domains, each of which is scored independently: • scope and purpose described • stakeholder (citizen/patient and professional) involvement • rigour of development (with evidence syntheses used as an input, a robust recommendations-development process, and recommendations linked to the supporting evidence) • clarity of presentation • applicability • editorial independence (in relation to funder and panel members' conflicts of interest)
		Grading of Recommendations, Assessment, Development and Evaluations (GRADE; bit.ly/3C9pMrx) for assessing the strength of recommendations, which uses four key considerations: • balance between desirable and undesirable outcomes (trade-offs), taking into account best estimates of the magnitude of effects on desirable and undesirable outcomes, and the importance of those outcomes (estimated typical values and preferences) • confidence in the magnitude of estimates of effects of the interventions on important outcomes (see GRADE in a previous row) • confidence in values and preferences and their variability resource use

Types of evidence for which quality-assessment tools don't yet exist

	Modeling	No widely accepted tool yet exists for most types of models, however, there are some general questions that can be asked about models (much like those listed as part of the Philips checklist above), such as: • structure of the model (e.g., explicit rationale, justified assumptions, and appropriate time horizon) • data used (e.g., baseline probabilities from observational studies, intervention effects from a range of sources*, and assessments of four types of uncertainty, namely the structure of the model, the methodological steps followed, the heterogeneity in the population studied, and the parameters used) • consistency (internal and external) • availability of the software or tool so that it can be assessed by others *One of the challenges with COVID-19 was that study designs typically used to capture intervention effects, such as randomized-controlled trials, were ethically or logistically difficult and/or took time to complete, so other study designs needed to be used and expert opinion needed to be sought (and there are approaches that enable this to be done in a way that is systematic and transparent, such as SHELF – see bit.ly/30nteC4)

Approaches used with certain types of evidence for which quality-assessment tools don't yet exist

	Artificial intelligence	No widely accepted tool yet exists