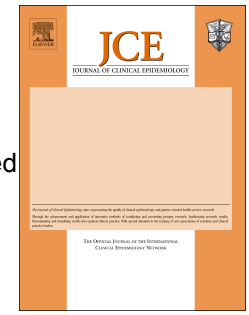


GRADE Guidelines 30: The GRADE Approach to Assessing the Certainty of Modelled Evidence - an Overview in the Context of Health Decision-making

Jan L. Brozek, Carlos Canelo-Aybar, Elie A. Akl, James M. Bowen, John Bucher, Weihsueh A. Chiu, Mark Cronin, Benjamin Djulbegovic, Maicon Falavigna, Gordon H. Guyatt, Ami A. Gordon, Michele Hilton Boon, Raymond C.W. Hutubessy, Manuela A. Joore, Vittal Katikireddi, Judy LaKind, Miranda Langendam, Veena Manja, Kristen Magnuson, Alexander G. Mathioudakis, Joerg Meerpohl, Dominik Mertz, Roman Mezencev, Rebecca Morgan, Gian Paolo Morgano, Reem Mustafa, Martin O'Flaherty, Grace Patlewicz, John J. Riva, Margarita Posso, Andrew Rooney, Paul M. Schlosser, Lisa Schwartz, Ian Shemilt, Jean-Eric Tarride, Kristina A. Thayer, Katya Tsaioun, Luke Vale, John Wambaugh, Jessica Wignall, Ashley Williams, Feng Xie, Yuan Zhang, Holger J. Schünemann, for the GRADE Working Group

PII: S0895-4356(20)31103-3

DOI: https://doi.org/10.1016/j.jclinepi.2020.09.018

Reference: JCE 10283

To appear in: *Journal of Clinical Epidemiology*

Received Date: 21 February 2020

Revised Date: 8 September 2020

Accepted Date: 17 September 2020

Please cite this article as: Brozek JL, Canelo-Aybar C, Akl EA, Bowen JM, Bucher J, Chiu WA, Cronin M, Djulbegovic B, Falavigna M, Guyatt GH, Gordon AA, Boon MH, Hutubessy RCW, Joore MA, Katikireddi V, LaKind J, Langendam M, Manja V, Magnuson K, Mathioudakis AG, Meerpohl J, Mertz D, Mezencev R, Morgan R, Morgano GP, Mustafa R, O'Flaherty M, Patlewicz G, Riva JJ, Posso M, Rooney A, Schlosser PM, Schwartz L, Shemilt I, Tarride J-E, Thayer KA, Tsaioun K, Vale L, Wambaugh J, Wignall J, Williams A, Xie F, Zhang Y, Schünemann HJ, for the GRADE Working Group, GRADE Guidelines 30: The GRADE Approach to Assessing the Certainty of Modelled Evidence - an Overview in the Context of Health Decision-making, *Journal of Clinical Epidemiology* (2020), doi: https://doi.org/10.1016/j.jclinepi.2020.09.018.

Author Statement

All authors analysed and interpreted the data. Jan Brozek and Carlos Canelo-Aybar wrote the first version of the paper. All authors of this paper have read and approved the final version submitted.

GRADE Guidelines 30: The GRADE Approach to Assessing the Certainty of Modelled Evidence - an Overview in the Context of Health Decision-making

Jan L. Brozek[# a,b,c], Carlos Canelo-Aybar[# d, e], Elie A. Akl[f], James M. Bowen[a,g], John Bucher[h], Weihsueh A. Chiu[i], Mark Cronin[j], Benjamin Djulbegovic[k], Maicon Falavigna[l], Gordon H. Guyatt[a,b,c], Ami A. Gordon[m], Michele Hilton Boon[n], Raymond C. W. Hutubessy[o], Manuela A. Joore[p], Vittal Katikireddi[n], Judy LaKind[q,r], Miranda Langendam[s], Veena Manja[a,t,u], Kristen Magnuson[m], Alexander G. Mathioudakis[v], Joerg Meerpohl[w,x], Dominik Mertz[a], Roman Mezencev[y], Rebecca Morgan[a], Gian Paolo Morgano[a,c], Reem Mustafa[a,z], Martin O'Flaherty[aa], Grace Patlewicz[ab], John J. Riva[c,ac], Margarita Posso[e], Andrew Rooney[h], Paul M. Schlosser[y], Lisa Schwartz[a], Ian Shemilt[ad], Jean-Eric Tarride[a,ae], Kristina A. Thayer[u], Katya Tsaioun[af], Luke Vale[ag], John Wambaugh[ab], Jessica Wignall[m], Ashley Williams[m], Feng Xie[a], Yuan Zhang[a,ah], Holger J. Schünemann[a,b,c], for the GRADE Working Group

Co-first author

Affiliations:

[a] Department of Health Research Methods, Evidence, and Impact, McMaster University, Hamilton, Ontario, Canada

[b] Department of Medicine, McMaster University, Hamilton, Ontario, Canada

[c] McMaster GRADE Centre & Michael DeGroote Cochrane Canada Centre, McMaster University, Hamilton, Ontario, Canada

[d] Department of Paediatrics, Obstetrics and Gynaecology, Preventive Medicine, and Public Health. PhD Programme in Methodology of Biomedical Research and Public Health. Universitat Autònoma de Barcelona, Bellaterra, Spain.

[e] Iberoamerican Cochrane Center, Biomedical Research Institute (IIB Sant Pau-CIBERESP), Barcelona, Spain

[f] Department of Internal Medicine, American University of Beirut, Beirut, Lebanon

[g] Toronto Health Economics and Technology Assessment (THETA) Collaborative, Toronto, Ontario, Canada

[h] National Toxicology Program, National Institute of Environmental Health Sciences, Durham, North Carolina, USA

[i] Texas A&M University, College Station, Texas, USA
[j] Liverpool John Moores University, Liverpool, UK
[k] Center for Evidence-Based Medicine and Health Outcome Research, Morsani College of Medicine, University of South Florida, Tampa, Florida, USA
[l] Institute for Education and Research, Hospital Moinhos de Vento, Porto Alegre, Rio Grande do Sul, Brazil
[m] ICF International, Durham, North Carolina, USA
[n] Institute of Health & Wellbeing, University of Glasgow, Glasgow, UK
[o] World Health Organization, Geneva, Switzerland
[p] Clinical Epidemiology and Medical Technology Assessment, Maastricht University Medical Centre+, Maastricht, the Netherlands
[q] LaKind Associates, LLC, Catonsville, Maryland, USA
[r] Department of Epidemiology and Public Health, University of Maryland School of Medicine, Baltimore, Maryland, USA
[s] Department of Clinical Epidemiology, Biostatistics, and Bioinformatics, Academic Medical Center, University of Amsterdam, Amsterdam, the Netherlands
[t] Department of Surgery, University of California Davis, Sacramento, California, USA
[u] Department of Medicine, Department of Veterans Affairs, Northern California Health Care System, Mather, California, USA
[v] Division of Infection, Immunity and Respiratory Medicine, University Hospital of South Manchester, University of Manchester, Manchester, UK
[w] Institute for Evidence in Medicine, Medical Center, University of Freiburg, Freiburg-am-Breisgau, Germany
[x] Cochrane Germany, Freiburg-am-Breisgau, Germany
[y] National Center for Environmental Assessment, U.S. Environmental Protection Agency, Washington, D.C., District of Columbia, USA
[z] Department of Medicine, University of Kansas Medical Center, Kansas City, Kansas, USA
[aa] Institute of Population Health Sciences, University of Liverpool, Liverpool, UK
[ab] National Center for Computational Toxicology, U.S. Environmental Protection Agency, Durham, North Carolina, USA
[ac] Department of Family Medicine, McMaster University, Hamilton, Ontario, Canada
[ad] EPPI-Centre, Institute of Education, University College London, London, UK
[ae] Programs for Assessment of Technology in Health, McMaster University, Hamilton, Ontario, Canada
[af] Evidence-Based Toxicology Collaboration, Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland, USA
[ag] Health Economics Group, Institute of Health and Society, Newcastle University, Newcastle upon Tyne, UK

[ah] Health Quality Ontario, Toronto, Ontario, Canada

Corresponding author:
Jan Brozek
McMaster University
Health Sciences Centre, Area 2C
1280 Main Street West
Hamilton, ON L8S 4K1, Canada

Abstract

Objectives:

To present the Grading of Recommendations Assessment, Development, and Evaluation (GRADE) conceptual approach to the assessment of certainty of evidence from modelling studies (i.e. certainty associated with model outputs).

Study Design and Setting:

Expert consultations and, an international multi-disciplinary workshop informed development of a conceptual approach to assessing the certainty of evidence from models within the context of systematic reviews, health technology assessments, and health care decisions. The discussions also clarified selected concepts and terminology used in the GRADE approach and by the modelling community. Feedback from experts in a broad range of modelling and health care disciplines addressed the content validity of the approach.

Results:

Workshop participants agreed, that the domains determining the certainty of evidence previously identified in the GRADE approach (risk of bias, indirectness, inconsistency, imprecision, reporting bias, magnitude of an effect, dose-response relation, and the direction of residual confounding) also apply when of assessing the certainty of evidence from models. The assessment depends on the nature of model inputs and the model itself and on whether one is evaluating evidence from a single model or multiple models. We propose a framework for selecting the best available evidence from models: 1) developing *de novo* a model specific to the situation of interest, 2) identifying an existing model the outputs of which provide the highest certainty evidence for the situation of interest, either "off the shelf" or after adaptation, and 3) using outputs from multiple models. We also present a summary of preferred terminology to facilitate communication among modelling and health care disciplines.

Conclusions:

This conceptual GRADE approach provides a framework for using evidence from models in health decision making and the assessment of certainty of evidence from a model or models. The GRADE Working Group and the modelling community are currently developing the detailed methods and related guidance for assessing specific domains determining the certainty of evidence from models across health care-related disciplines (e.g. therapeutic decision-making, toxicology, environmental health, health economics).

Introduction

When direct evidence to inform health decisions is not available or not feasible to **measure** (e.g. long-term effects of interventions or when studies in certain populations are perceived as unethical), modelling studies may be used to **predict** that "evidence" and inform decision-making.[1, 2] Health decision makers arguably face many more questions than can be reasonably answered with studies that directly measure the outcomes. Modelling studies, therefore, are increasingly used to predict disease dynamics and burden, the likelihood that an exposure represents a health hazard, the impact of interventions on health benefits and harms, or the economic efficiency of health interventions, among others [1]. Irrespective of the modelling discipline, decision makers need to know the best **estimates** of the modelled outcomes and how much **confidence** they may have in each estimate.[3] Knowing to what extent one can trust the outputs of a model is necessary when using them to support health decisions [4].

Although a number of guidance documents on how to assess the trustworthiness of estimates obtained from models in several health fields have been previously published [5-16], they are limited by failing to distinguish methodological rigor from completeness of reporting, and by failing to clear distinguish among various components affecting the trustworthiness of model outputs. In particular they lack clarity regarding sources of uncertainty that may arise from **model inputs** and from the uncertainty about a **model itself**. Modellers and those using results from models should assess the credibility of both.[4]

Authors have attempted to develop tools to assess model credibility, but many addressed only selected aspects, such as statistical reproducibility of data, the quality of reporting[17], or a combination of reporting with aspects of good modelling practices[7, 18-21]. Many tools also do not provide sufficiently detailed guidance on how to apply individual domains or criteria. There is therefore a need for further development and validation of such tools in specific disciplines. Sufficiently detailed guidance for making and reporting these assessments is also necessary.

Models predict outcomes based on model inputs – previous observations, knowledge and assumptions about the situation being modelled. Thus, when developing new models or assessing whether an existing model has been optimally developed, one should specify *a priori* the most appropriate and relevant data sources to inform different parameters required for the model. These may be either (seldom) a single study that provides the most direct information for the situation being modelled or (more commonly) a systematic review of multiple studies that identify all relevant sources of data. The risk of bias, directness and consistency of input data, precision of these estimates, and other domains specified in the Grading of Recommendations Assessment,

Development, and Evaluation (GRADE) approach determine the certainty of each of the model inputs.[22-28]

When assessing the evidence generated, various disciplines in health care and related areas that use modelling face similar challenges may benefit from shared solutions. Table 1 presents examples of selected models used in health-related disciplines in Table 1. Building on the existing GRADE approach, werefined and expand guidance regarding assessment of the certainty of model outputs. We formed a GRADE project group comprised of individuals with expertise in developing models and using model results in health-related disciplines, to create a unified framework for assessing the certainty of model outputs in the context of systematic reviews [29], health technology assessments, health care guidelines, and other health decision-making. In this article, we outline the proposed conceptual approach and clarify key terminology (Table 2). The target audience for this article includes researchers who develop models and those who use models to inform health care-related decisions.

What we mean by a model

Authors have used the term *model* to describe a variety of different concepts [2] and suggested several broader or narrower definitions [6, 30], so even modellers in the relatively narrow context of health sciences can differ in their views regarding what constitutes a model. Models vary in their structure and degree of complexity. A very simple model might be an equation estimating a variable not directly measured, such as the absolute effect of an intervention estimated as the product of the intervention's relative effect and the assumed baseline risk in a defined population (risk difference equals relative risk reduction multiplied by an assumed baseline risk). On the other end of the spectrum, elaborate mathematical models, such as system dynamics models (e.g. infectious disease transmission) may contain dozens of sophisticated equations that require considerable computing power to solve.

By their nature, such models only *resemble* the phenomena being modelled – i.e. specific parts of the world that are interesting in the context of a particular decision – with necessary approximations and simplifications, and to the extent that one actually knows and understands the underlying mechanisms.[1] Given the complexity of the world, decision-makers often rely on some sort of a model to answer health-related questions.

In this article, we focus on quantitative mathematical models defined as "mathematical framework representing variables and their interrelationships to describe observed phenomena or predict future events"[30] used in health-related disciplines for decision-making (Table 1). These may be models of systems representing causal mechanisms (aka mechanistic models), models

predicting outcomes from input data (aka empirical models), and models combining mechanistic with empirical approaches (aka hybrid models). We do not consider here statistical models used to estimate the associations between measured variables (e.g. proportional hazards models or models used for meta-analysis).

The GRADE approach

The GRADE working group was established in the year 2000 and continues as a community of people striving to create systematic, and transparent frameworks for assessing and communicating the certainty of the available evidence used in making decisions in healthcare and health-related disciplines.[31] The GRADE Working Group now includes over 600 active members from 40 countries and serves as a think tank for advancing evidence-based decision-making in multiple health-related disciplines (www.gradeworkinggroup.org). GRADE is widely used internationally by over 110 organizations to address topics related to clinical medicine, public health, coverage decisions, health policy, and environmental health.

The GRADE framework uses concepts familiar to health scientists, grouping specific items to evaluate the certainty of evidence in conceptually coherent domains. Specific approaches to the concepts may differ depending on the nature of the body of evidence (Table 2). GRADE domains include concepts such as risk of bias[28], directness of information [24], precision of an estimate[23], consistency of estimates across studies[25], risk of bias related to selective reporting[26], strength of the association, presence of a dose-response gradient, and the presence of plausible residual confounding that can increase confidence in estimated effects[27].

The general GRADE approach is applicable irrespective of health discipline. It has been applied to rating the certainty of evidence for management interventions, health care related tests and strategies [32, 33], prognostic information[34], evidence from animal studies[35], use of resources and cost-effectiveness evaluations[36], and values and preferences[37, 38]. Although the GRADE Working Group has begun to address certainty of modelled evidence in the context of test-treatment strategies[39], health care resource use and costs[36], and environmental health[40], more detailed guidance is needed for complex models such as those used in infectious diseases, health economics, public health, and decision analysis.

Methods

On May 15 and 16, 2017, health scientists participated in a GRADE modelling project group workshop in Hamilton, Ontario, Canada, to initiate a collaboration in developing common

principles for the application of the GRADE assessment of certainty of evidence to modelled outputs. The National Toxicology Program of the Department of Health and Human Services in the USA and the MacGRADE Center in the Department of Health Research Methods, Evidence, and Impact at McMaster University sponsored the workshop which was co-organized by MacGRADE Center and ICF International.

Workshop participants were selected to ensure a broad representation of all modelling related fields (Appendix). Participants had expertise in modelling in the context of clinical practice guidelines, public health, environmental health, dose-response modelling, physiologically based pharmacokinetic (PBPK) modelling, environmental chemistry, physical/chemical property prediction, evidence integration, infectious disease, computational toxicology, exposure modelling, prognostic modelling, diagnostic modelling, cost effectiveness modelling, biostatistics, and health ethics.

Leading up to the workshop, we held three webinars to introduce participants to the GRADE approach. Several workshop participants (VM, KT, JB, AR, JW, JLB, HJS) collected and summarized findings from literature and the survey of experts as background material that provided a starting point for discussion. The materials included collected terminology representing common concepts across multiple disciplines that relate to evaluating modelled evidence, and a draft framework for evaluating modelled evidence. Participants addressed specific tasks in small groups and large group discussion sessions and agreed on key principles both during the workshop and through written documents.

Results

Terminology

Workshop participants agreed on the importance of clarifying terminology to facilitate communication among modellers, researchers, and users of model outputs from different disciplines. Modelling approaches evolved somewhat independently, resulting in different terms being used to describe the same or very similar concepts or the same term being used to describe different concepts. For instance, the concept of extrapolating from the available data to the context of interest has been referred to as directness, applicability, generalizability, relevance, or external validity. The lack of standardized terminology leads to confusion and hinders effective communication and collaboration among modellers and users of models.

Overcoming these obstacles would require clarifying the definitions of concepts and agreeing on terminology across disciplines. Realizing that this involves changing established customary use of terms in several disciplines, workshop participants suggested accepting the use of alternative terminology while always being clear about the preferred terms to be used and the underlying concept to which it refers (Table 2). Experts attending a World Health Organization's consultation have very recently suggested a more extensive set of terms [41]. To facilitate future communication, participants of this workshop will further collaborate to build a comprehensive glossary of terminology related to modelling.

Outline of an approach to using model outputs for decision making

Workshop participants suggested an approach to incorporate model outputs in health-related decision making (Figure 1). In this article we describe only the general outline of the suggested approach – in subsequent articles we will discuss the details of the approach and provide more specific guidance on its application to different disciplines and contexts.

Researchers should start by **conceptualizing the problem and the ideal target model** that would best represent the actual phenomenon or decision problem they are considering [13]. This conceptualization would either guide the development of a new model or serve as a reference against which existing models could be compared. The ideal target model should reflect: 1) the relevant population (e.g., patients receiving some diagnostic procedure or exposed to some hazardous substance), 2) the exposures or health interventions being considered, 3) the outcomes of interest in that context, and 4) their relationships. [42]. Conceptualizing the model will also reduce the risk of intentional or unintentional development of data-driven models, in which inputs and structure would be determined only by what is feasible to develop given the available data at hand.

Participants identified 3 options in which users may incorporate model outputs in health decision-making (Figure 1):

1. **Develop a model de novo designed specifically to answer the very question at hand.** Workshop participants agreed that in an ideal situation such an approach would almost always be the most appropriate. Following this approach, however, requires suitable skills, ample resources, and time being available. It also requires enough knowledge about the phenomenon being modelled to be able to tell whether or not the new model would have any advantage over already existing models.
2. **Search for an existing model describing the same or a very similar problem and use it** "*off-the-shelf*" or adapt it appropriately in order to answer the current question. In practice many researchers initially use this approach because of the above limitations of developing a new

model. However, it is often not possible to find an existing model that would be directly relevant to the problem at hand and/or it is not feasible to adapt an existing model when found. Any adaptation of a model requires availability of input data relevant for current problem, appropriate expertise and resources, and access to the original model. The latter is often not available (e.g. proprietary model or no longer maintained) or the structure of the original model is not being transparent enough to allow adaptation ("black-box").

3. **Use the results from multiple existing models** found in the literature [43]. This approach may be useful when a limited knowledge about the phenomenon being modelled makes it impossible to decide which of the available models is more relevant, or when many alternative models are relevant but use different input parameters. In such situations, one may be compelled to rely on the results of several models, because selection of the single, seemingly "best" model may provide incorrect estimates of outputs and lead to incorrect decisions.

Identifying existing models that are similar to the ideal target model often requires performing a scoping of the literature or a complete **systematic review** of potentially relevant models – a structured process following a standardized set of methods with a goal to identify and assess all available models that are accessible, transparently reported, and fulfil the pre-specified eligibility criteria based on the conceptual ideal target model. Some prefer the term **systematic survey** that differs from a systematic review in the initial intention to use the results: in systematic reviews the initial intention is to combine the results across studies either statistically through a meta-analysis or narratively summarizing their results when appropriate, whereas in a systematic survey the initial intention is to examine the various ways that an intervention or exposure has been modelled, to review the input evidence that has been used, and ultimately to identity a single model that fits the conceptual ideal target model the best or requires the least adaptation; only when one cannot identify a single such model will it be necessary to use the results of multiple existing models.

If a systematic search revealed one or more models meeting the eligibility criteria, then researchers would assess the certainty of outputs from each model. Depending on this assessment, researchers may be able to use the results of a single most direct and lowest risk of bias model "off-the-shelf" or proceed to adapt that model. If researchers failed to find an existing model that would be sufficiently direct and low risk of bias, then they would ideally develop their own model de novo.

Assessing the certainty of outputs from a single model

When researchers develop their own model or when they identify a single model that is considered sufficiently direct to the problem at hand, they should assess the certainty of its

outputs (i.e. evidence generated from that model). Note, that if a model estimates multiple outputs, researchers needs to assess the certainty of each output separately [23-28]. Workshop participants agreed that all GRADE domains are applicable to assess the certainty of model outputs, but further work is needed to identify examples and develop specific criteria to be assessed, which may differ depending on the model being used and/or situation being modelled.

Risk of bias in a single model

The risk of bias of model outputs (i.e. model outputs being systematically overestimated or underestimated) is determined by the credibility of a model itself and the certainty of evidence for each of model inputs.

The **credibility of a model**, also referred to as the quality of a model (Table 2) is influenced by its conceptualization, structure, calibration, validation, and other factors. Determinants of model credibility are likely to be specific to a modelling discipline (e.g., health economic models have different determinants of their credibility than PBPK models). There are some discipline-specific guidelines or checklists developed for the assessment of credibility of a model and other factors affecting the certainty of model outputs such as the framework to assess adherence to good practice guidelines in decision-analytic modelling [18], the questionnaire to assess relevance and credibility of modelling studies [18, 44, 45], good research practices for modelling in health technology assessment [5, 6, 8, 9, 12-14], the approaches to assessing uncertainty in read-across [46], and the quantitative structure-activity relationships [47] in predictive toxicology. Workshop participants agreed that there is a need for comprehensive tools developed specifically to assess credibility of various types of models in different modelling disciplines.

The **certainty of evidence in each of the model inputs** is another critical determinant of the risk of bias in a model. A model has several types of input data – bodies of evidence used to populate a model (Table 2). When researchers develop their model *de novo*, in order to minimize the risk of bias they need to specify those input parameters to which the model outputs are the most sensitive. For instance, in economic models these key parameters may include health effects, resource use, utility values, and baseline risks of outcomes. Model inputs should reflect the entire body of relevant evidence satisfying clear pre-specified criteria rather than an arbitrarily selected evidence that is based on convenience (“any available evidence”) or picked in any other non-systematic way (e.g., “first evidence found” – single studies that researchers happen to know about or are the first hits in a database search).

The appropriate approach will depend on the type of data and may require performing a systematic review of evidence on each important or crucial input variable [48-50]. Some inputs

may have a very narrow inclusion criteria and therefore evidence from single epidemiological survey or population surveillance may provide all relevant data for the population of interest (e.g. baseline population incidence or prevalence).

The certainty of evidence for each input needs to be assessed following the established GRADE approach specific to that type of evidence (e.g. estimates of intervention effects or baseline risk of outcomes)[22, 32, 34, 37]. Following the logic of the GRADE approach that the overall certainty of evidence cannot be higher than the lowest certainty for any body of evidence that is critical for a decision [51], the overall rating of certainty of evidence across model inputs should be limited by the lowest certainty rating for any body of evidence (in this case input data) to which the model output(s) was proved sensitive.

Application of this approach requires a priori consideration of likely critical and/or important inputs when specifying the *conceptual ideal target model* and the examination of the results of *back-end* sensitivity analyses. It further requires deciding how to judge whether results are or are not sensitive to alternative input parameters. Authors have described several methods to identify the most influential parameters including global sensitivity analysis to obtain "parameter importance measures" (i.e. information based measures) [52]; or alternatively by varying one parameter at a time and assessing their influence in "base case" outputs [52] For example, in a model-based economic evaluation one might be looking for the influence of sensitivity analysis on cost-effectiveness ratios at a specified willingness-to-pay threshold.

Indirectness in a single model

By directness or relevance, we mean the extent to which model outputs directly represent the phenomenon being modelled. To evaluate the relevance of a model, one needs to compare it against the conceptual ideal target model. When there are concerns about the directness of the model or there is limited understanding of the system being modelled making it difficult to assess directness, then one may have lower confidence in model outputs.

Determining the directness of model outputs includes assessing to what extent the modelled population, the assumed interventions and comparators, the time horizon, the analytic perspective, as well as the outcomes being modelled reflect those that are current interest. For instance, if the question is about the risk of birth defects in children of mothers chronically exposed to a certain substance, there may be concerns about the directness of the evidence if the model assumed short-term exposure, the route of exposure was different, or the effects of exposure to a similar but not the same substance were measured.

Assessing indirectness in a single model also requires evaluating two separate sources of indirectness:

1. indirectness of input data with respect to the ideal target model´s inputs.

2. indirectness of model outputs with respect to the decision problem at hand.

This conceptual distinction is important because, although they are interrelated, one needs to address each type of indirectness separately. Even if the outputs might be direct to the problem of interest, the final assessment should consider if the inputs used were also direct for the target model.

Using an existing model has potential limitations: its inputs might have been direct for the decision problem addressed by its developers but are not direct with respect to the problem currently at hand. In this context, sensitivity analysis can help to assess to what extent model outputs are robust to the changes in input data or assumptions used in model development.

Inconsistency in a single model

A single model may yield inconsistent outputs owing to unexplained variability in the results of individual studies informing the pooled estimates of input variables. For instance, when developing a health economic model, a systematic review may yield several credible, but discrepant, utility estimates in the population of interest. If there is no plausible explanation for that difference in utility estimates, outputs of a model based on those inputs may also be qualitatively inconsistent. Again, sensitivity analysis may help to make a judgment to what extent such inconsistency of model inputs would translate into a meaningful inconsistency in model outputs with respect to the decision problem at hand.

Imprecision in a single model

Sensitivity analysis characterizes the response of model outputs to parameter variation, and helps to determine the robustness of model´s qualitative conclusions [52, 53]. The overall certainty of model outputs may also be lower when the outputs are estimated imprecisely. For quantitative outputs one should examine not only the point estimate (e.g., average predicted event) but also the variability of that estimate (e.g., results of the probabilistic sensitivity analysis based in the distribution of the input parameters). It is essential that a report from a modelling study always includes information about output variability. Further guidance on how to assess imprecision in model outputs will need to take into account if the conclusions change according to that specific parameter. In some disciplines, for instance in environmental health, model inputs are frequently qualitative. Users of such models may assess “adequacy” of the data, i.e. the degree of “richness” and quantity of data supporting particular outputs of a model.

Risk of publication bias in the context of a single model

The risk of publication bias, also known as "reporting bias", "non-reporting bias", or "bias owing to missing results", as it is currently called in the Cochrane Handbook [54], is the likelihood that relevant models have been constructed but were not published or otherwise made publicly available. Risk of publication bias may not be relevant when assessing the certainty of outputs of a single model constructed de novo. However, when one intends to reuse an existing model but is aware or strongly suspects that similar models had been developed but are not available, then one may be inclined to think that their outputs might have systematically differed from the model that is available. In such a case, one may have lower confidence in the outputs of the identified model if there is no reasonable explanation for the inability to obtain those other models.

Domains that increase the certainty of outputs from a single model

The GRADE approach to rating the certainty of evidence recognized three situations when the certainty of evidence can increase: large magnitude of an estimated effect, presence of a dose-response gradient in an estimated effect, and an opposite direction of plausible residual confounding.[27] Workshop participants agreed that presence of a **dose-response gradient** in model outputs may be applicable in some modelling disciplines (e.g., environmental health). Similarly, whether or not a **large magnitude of an effect** in model outputs increases the certainty of the evidence may depend on the modelling discipline. The **effect of an opposite direction of plausible residual confounding** seems theoretically also applicable in assessing the certainty of model outputs (i.e. a conservative model not incorporating input data parameter in favour of an intervention but still finding favorable outputs) but an actual example of this phenomenon in modelling studies is still under discussion.

Assessing the certainty of outputs across multiple models

Not infrequently, particularly in disciplines relying on mechanistic models, the current knowledge about the real system being modelled is very limited precluding the ability to determine which of the available existing models generates higher certainty outputs. Therefore, it may be necessary to rely on the results across multiple models. Other examples include using multiple models when no model was developed for the population directly of interest (e.g. the European Breast Cancer Guideline for Screening and Diagnosis relied on a systematic review of modelling studies that compared different mammography screening intervals [55]) or when multiple models of the same situation exist but vary in structure, complexity, and parameter choices (e.g. HIV Modelling Consortium compared several different mathematical models simulating the same antiretroviral

493 therapy program and found that all models predicted that the program has the potential to
494 reduce new HIV infections in the population [56]).
495
496 When researchers choose or are compelled to include outputs from several existing models, they
497 should assess the certainty of outputs across all included models. This assessment may be more
498 complex than for single models and single bodies of evidence. The feasibility of GRADE's guidance
499 to judge the certainty of evidence lies in the availability of accepted methods for assessing most
500 bodies of evidence from experimental to observational studies. However, the methods for
501 systematic reviews of modelling studies are less well-established, some stages of the process are
502 more complex, the number of highly skilled individuals with experience in such systematic reviews
503 is far lower, and there is larger variability in the results [57]. Additionally, researchers must be
504 careful to avoid "double counting" the same model as if it were multiple models. For instance, the
505 same model (i.e. same structure and assumptions) may have been used in several modelling
506 studies, in which investigators relied on different inputs. When facing this scenario, researchers
507 may need to decide which of the inputs are the most direct to their particular question and
508 include in only this model in the review.
509
510 *Risk of bias across multiple models*
511
512 The assessment of risk of bias across models involves an assessment of the risk of bias in each
513 individual model (see above discussion of risk of bias in single model) and subsequently making a
514 judgement about the overall risk of bias across all included models. Specific methods for
515 operationalizing this integration remain to be developed.
516
517 *Indirectness across multiple models*
518
519 As for the risk of bias, researchers need to assess indirectness of outputs initially for each of
520 included models and then integrate the judgements across models. Likewise, specific methods for
521 operationalizing this integration still remain to be developed. During this assessment researchers
522 may find some models too indirect to be informative for their current question and decide to
523 exclude them from further consideration. However, the criteria to determine which models are
524 too indirect should be developed a priori, before the search for the models is performed and their
525 results are known.
526
527 *Imprecision across multiple models*
528
529 The overall certainty of model outputs may also be lower when model outputs are not estimated
530 precisely. If researchers attempt a quantitative synthesis of outputs across models, they will

report the range of estimates and variability of that estimates. When researchers choose to perform only a qualitative summary of the results across models, it is desirable that they report some estimate of variability in the outputs of individual models and an assessment of how severe the variability is (e.g. range of estimated effects).

Inconsistency of outputs across multiple models

The assessment of inconsistency should focus on unexplained differences across model outputs for a given outcome. If multiple existing models addressing the same issue produce considerably different outputs or reach contrasting conclusions, then careful comparison of the models may lead to a deeper understanding of the factors that drive outputs and conclusions. Ideally, the different modelling groups that developed relevant models would come together to explore the importance of differences in the type and structure of their models, and of the data used as model inputs.

Invariably there will be some differences among the estimates from different models. Researchers will need to assess whether or not these differences are important, i.e. whether they would lead to different conclusions. If the differences are important but can be explained by model structure, model inputs, the certainty of the evidence of the input parameters or other relevant reasons, one may present the evidence separately for the relevant subgroups. If differences are important, but cannot be clearly explained, the certainty of model outputs may be lower.

Risk of publication bias across multiple models

The assessment is similar to that of the risk of publication bias in the context of a single model.

Domains that increase the certainty of outputs across multiple models

All considerations are the same to those in the context of a single model.

Discussion

The goal of the GRADE project group on modelling is to provide concepts and operationalization of how to rate the certainty of evidence in model outputs. This article provides an overview of the conclusions of the project group. This work is important because there is a growing need and availability of modelled information resulting from a steadily increasing knowledge of the complexity of the structure and interactions in our environment, and computational power to

construct and run models. Users of evidence obtained from modelling studies need to know how much trust they may have in model outputs. There is a need to improve the methods of constructing models and to develop methods for assessing the certainty in model outputs. In this article we have attempted to clarify the most important concepts related to developing and using model outputs to inform health-related decision-making. Our preliminary work identified confusion about terminology, lack of clarity of what is a model, and need for methods to assess certainty in model outputs as priorities to be addressed in order to improve the use of evidence from modelling studies.

In some situations, decision-makers might be better off developing a new model specifically designed to answer their current question. However, we suggest that it is not always feasible to develop a new model or that developing a new model might not be any better than using already existing models, when the knowledge of the real life system to be modelled is limited precluding the ability to choose one model that would be better than any other. Thus, sometimes it may be necessary or more appropriate to use one or multiple existing models depending on their availability, credibility, and relevance to the decision-making context. The assessment of the certainty of model outputs will be conceptually similar when a new model is constructed, or one existing model is used. The main difference between the latter two approaches is the availability of information to perform a detailed assessment. That is, information for one's own model may be easily accessible, but information required to assess someone else's model will often be more difficult to obtain. Assessment of the certainty evidence across models can build on existing GRADE domains but requires different operationalization.

Because it builds on an existing, widely used framework that includes a systematic and transparent evaluation process, modelling disciplines' adoption of the GRADE approach and further development of methods to assess the certainty of model outputs may be beneficial for health decision making. Systematic approaches improve rigor of research, reducing the risk of error and its potential consequences; transparency of the approach increases its trustworthiness. There may be additional benefits related to other aspects of the broader GRADE approach, for instance a potential to reduce unnecessary complexity and workload in modelling by careful consideration of the most direct evidence as model inputs. This may allow, for instance, optimization of the use of different streams of evidence as model inputs. Frequently, authors introduce unnecessary complexity by considering multiple measures of the same outcome when focus could be on the most direct outcome measure.

The GRADE working group will continue developing methods and guidance for using model outputs in health-related decision-making. In subsequent articles we will provide more detailed guidance about choosing the "best" model when multiple models are found, using multiple

606 models, integrating the certainty of evidence from various bodies of evidence with credibility of
607 the model and arriving at the overall certainty in model outputs, how to assess the credibility of
608 various types of models themselves, and further clarification of terminology. In the future we aim
609 to develop and publish the detailed guidance for assessing certainty of evidence from models, the
610 specific guidance for the use of modelling across health care-related disciplines (e.g. toxicology,
611 environmental health or health economics), validation of the approach, and accompanying
612 training materials and examples.
613

614 Acknowledgments

615

616 AR was supported by the National Institutes of Health, National Institute of Environmental Health
617 Sciences.

618 **Table 1.** Examples of modelling methods in health-related disciplines (not comprehensive)*
619

Decision analysis models	Structured model representing health care pathways examining effects of an intervention on outcomes of interest.
	Types ▪ Decision tree models ▪ State transition models o Markov cohort simulation o Individual based microsimulation (first-order Monte Carlo) ▪ Discrete event simulation ▪ Dynamic transmission models ▪ Agent based models
	Examples ▪ Estimation of long-term benefits and harms outcomes from complex intervention, e.g. minimum unit pricing of alcohol ▪ Estimation of benefits and harms of population mammography screening based in microsimulation model, e.g. Wisconsin model from CISNET collaboration[58] ▪ Susceptible-Infectious-Recovery transmission dynamic model to assess effectiveness of lockdown during the SARS-CoV-2 pandemic[59]
Pharmacology and toxicology models	Computational models developed to organize, analyse, simulate, visualize or predict toxicological and ecotoxicological effects of chemicals. In some cases, these models are used to estimate the toxicity of a substance even before it has been synthesized.
	Types ▪ Structural alerts and rule-based models ▪ Read-Across ▪ Dose response and Time response ▪ Toxicokinetic (TK) and toxicodynamic(TD) ▪ Uncertainty factors ▪ Quantitative structure activity relationship (QSAR) ▪ Biomarker-based toxicity models
	Examples • Structural alerts for mutagenicity and skin sensitisation • Read-across for complex endpoints such as chronic toxicity • Pharmacokinetic (PK) models to calculate concentrations of substances in organs, following a variety of exposures QSAR models for carcinogenicity • TGx-DDI biomarker to detect DNA damage-inducing agents
Environmental models	The EPA defined these models as: 'A simplification of reality that is constructed to gain insights into select attributes of a physical, biological, economic, or social system.' It involves the application of multidisciplinary knowledge to explain, explore and predict the Earth´s response to environmental change, and the interactions between human activities and natural processes.

	Classification (based on the CREM guidance document): • Human activity models • Natural systems process • Emission models • Fate and transport models • Exposure models • Human health effects models • Ecological effects models • Economic impact models • Noneconomic impact models
	Examples • Land use regression models • IH SkinPerm [60] • ConsExpo [61] • other exposure models [62]
Other	• HopScore: An Electronic Outcomes-Based Emergency Triage System [63] • Computational general equilibrium (CGE) models [64]
*Although not described in this classification simple calculations incorporating two or more pieces of evidence as for example the multiplication of a RR by the baseline risk to obtain the absolute risk difference of an intervention is a model, although pragmatic, with their respective assumptions.	

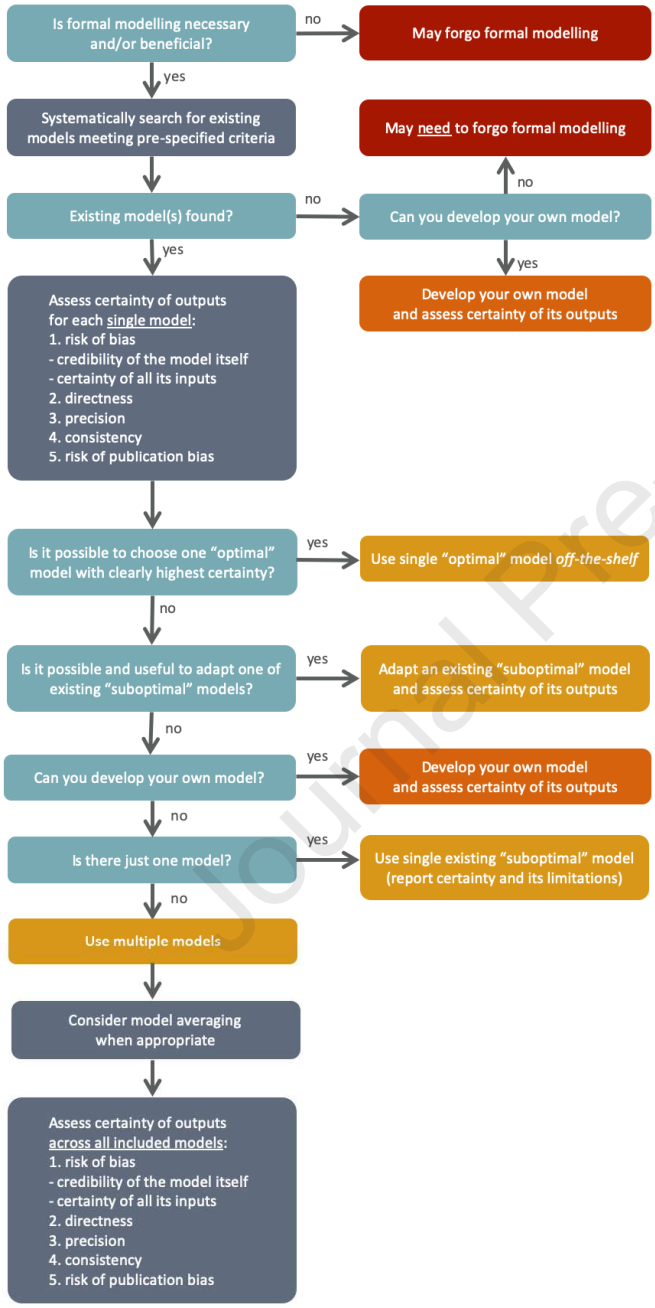
620
621
622 **Table 2.** Selected commonly used and potentially confusing terms used in the context of modelling
623 and the GRADE approach
624

Term	General definition
Sources of evidence (may come from in vitro or in vivo experiment or a mathematical model)	
Streams of evidence	Parallel **information about the same outcome that may have been obtained using different methods of estimating that outcome**. For instance, evidence of the increased risk for developing lung cancer in humans after an exposure to certain chemical compound may come from several streams of evidence: 1) mechanistic evidence – models of physiological mechanisms, 2) studies in animals – observations and experiments in animals from different phyla, classes, orders, families, genera, and species (e.g., bacteria, nematodes, insects, fish, mice, rats), and 3) studies in humans.
Bodies of evidence	Information about multiple different aspects around a decision about the best course of action. For instance, in order to decide whether or not a given diagnostic test should be used in some people, one needs to integrate the bodies of evidence about: the accuracy of the test, the prevalence of the conditions being suspected, the natural history of these conditions, the effects of potential treatments, values and preferences of affected individuals, cost, feasibility, etc.
Quality	

(may refer to many concepts, thus **alternative terms are preferred** to reduce confusion)	
Certainty of model outputs Alternative terms: ▪ certainty of modelled evidence ▪ quality of evidence ▪ quality of model output ▪ strength of evidence ▪ confidence in model outputs	In the context of health decision-making, the **certainty of evidence** (term preferred over "quality" in order to avoid confusion with the risk of bias in an individual study) reflects the extent to which one's confidence in an estimate of an effect is adequate to make a decision or a recommendation. Decisions are influenced not only by the best estimates of the expected desirable and undesirable consequences but also by one's confidence in these estimates. In the context of evidence syntheses of separate bodies of evidence (e.g., systematic reviews), the certainty of evidence reflects the extent of confidence that an estimate of effect is correct. For instance, the attributable national risk of cardiovascular mortality resulting from exposure to air pollution measured in selected cities. The GRADE Working Group published several articles explaining the concept in detail.[22-28, 65] Note that the phrase "confidence in an estimate of an effect" does not refer to statistical confidence intervals. Certainty of evidence is always assessed for the whole body of evidence rather than on a single study level (single studies are assessed for risk of bias and indirectness).
Certainty of model inputs Alternative term: ▪ quality of model inputs	Characteristics of data that are used to develop, train, or run the model, e.g., source of input values, their manipulation prior to input into a model, quality control, risk of bias in data, etc.
Credibility of a model Alternative terms: ▪ quality of a model ▪ risk of bias in a model ▪ validity of a model	To avoid confusion and keep with terminology used by modelling community[7] we suggest using the term *credibility* rather than *quality* of a model. The concept refers to the characteristics of a model itself – its design or execution – that affect the risk that the results may overestimate or underestimate the true effect. Various factors influence the overall credibility of a model, such as its structure, the analysis and the validation of the assumptions made during modelling.
Quality of reporting	Refers to how comprehensively and clearly model inputs, a model itself, and model outputs have been documented and described such that they can be critically evaluated and used for decision-making. Quality of reporting and quality of a model are separate concepts: a model with a low quality of reporting is not necessarily a low-quality model and vice versa.
Directness	
Directness of a model Alternative terms: ▪ relevance ▪ external validity ▪ applicability ▪ generalizability ▪ transferability ▪ translatability	By directness of a model we mean the extent to which the model represents the real-life situation being modelled which is dependent on how well the input data and the model structure reflect the scenario of interest. Directness is the term used in the GRADE approach, because each of the alternatives has been used usually in a narrower meaning.

626 * There may be either subtle or fundamental differences among some disciplines in how these
627 terms are being used; for the purposes of this article, these terms are generalized rather than
628 discipline specific.
629
630

631 **Figure 1.** The general approach to using modelled evidence and assessing its certainty in health-
632 related disciplines.
633



634
635

Appendix. List of workshop participants

Elie Akl (EA)– American University of Beirut, Lebanon
Jim Bowen (JMB)– McMaster University, Canada
Chris Brinkerhoff (CB)– US Environmental Protection Agency, USA
Jan Brozek (JLB)– McMaster University, Canada
John Bucher (JB)– US National Toxicology Program, USA
Carlos Canelo-Aybar (CCA)– Iberoamerican Cochrane Centre, Spain
Marcy Card (MC)– US Environmental Protection Agency, USA
Weihsueh A. Chiu (WCh)– Texas A&M University, USA
Mark Cronin (MC)– Liverpool John Moores University, UK
Tahira Devji (TD)– McMaster University, Canada
Ben Djulbegovic (BD)– University of South Florida, USA
Ken Eng (KE)– Public Health Agency of Canada
Gerald Gartlehner (GG)– Donau-Universität Krems, Austria
Gordon Guyatt (GGu)– McMaster University, Canada
Raymond Hutubessy (RH)– World Health Organization Initiative for Vaccine Research, Switzerland
Manuela Joore (MJ)– Maastricht University, the Netherlands
Richard Judson (RJ)– US Environmental Protection Agency, USA
S. Vittal Katikireddi (SK)– University of Glasgow, UK
Nicole Kleinstreuer (NK)– US National Toxicology Program, USA
Judy LaKind (JL)– University of Maryland, USA
Miranda Langendam (ML)– University of Amsterdam, the Netherlands
Zbyszek Leś (ZL)– Evidence Prime Inc., Canada
Veena Manja (VM)– McMaster University, Canada
Joerg Meerpohl (JM)– GRADE Center Freiburg, Cochrane Germany, University Medical Center Freiburg
Dominik Mertz (DM)– McMaster University, Canada
Roman Mezencev (RM)– US Environmental Protection Agency, USA
Rebecca Morgan (RMo)– McMaster University, Canada
Gian Paolo Morgano (GPM)– McMaster University, Canada
Reem Mustafa (RMu)– University of Kansas, USA
Bhash Naidoo (BN)– National Institute for Health and Clinical Excellence, UK
Martin O'Flaherty (MO)– Public Health and Policy, University of Liverpool, UK
Grace Patlewicz (GP)– US Environmental Protection Agency, USA
John Riva (JR)– McMaster University, Canada
Alan Sasso (AS)– US Environmental Protection Agency, USA
Paul Schlosser (PS)– US Environmental Protection Agency, USA

674 Holger Schünemann (HJS)– McMaster University, Canada
675 Lisa Schwartz (LS)– McMaster University, Canada
676 Ian Shemilt (IS)– University College London, UK
677 Marek Smieja (MS)– McMaster University, Canada
678 Ravi Subramaniam (RS)– US Environmental Protection Agency, USA
679 Jean-Eric Tarride (JT)– McMaster University, Canada
680 Kris Thayer (KAT)– US Environmental Protection Agency, USA
681 Katya Tsaioun (KT)– John Hopkins University, USA
682 Bernhard Ultsch (BU)– Robert Koch Institute, Germany
683 John Wambaugh (JW)– US Environmental Protection Agency, USA
684 Jessica Wignall (JWi)– ICF, USA
685 Ashley Williams (AW)– ICF, USA
686 Feng Xie (FX)– McMaster University, Canada
687

References

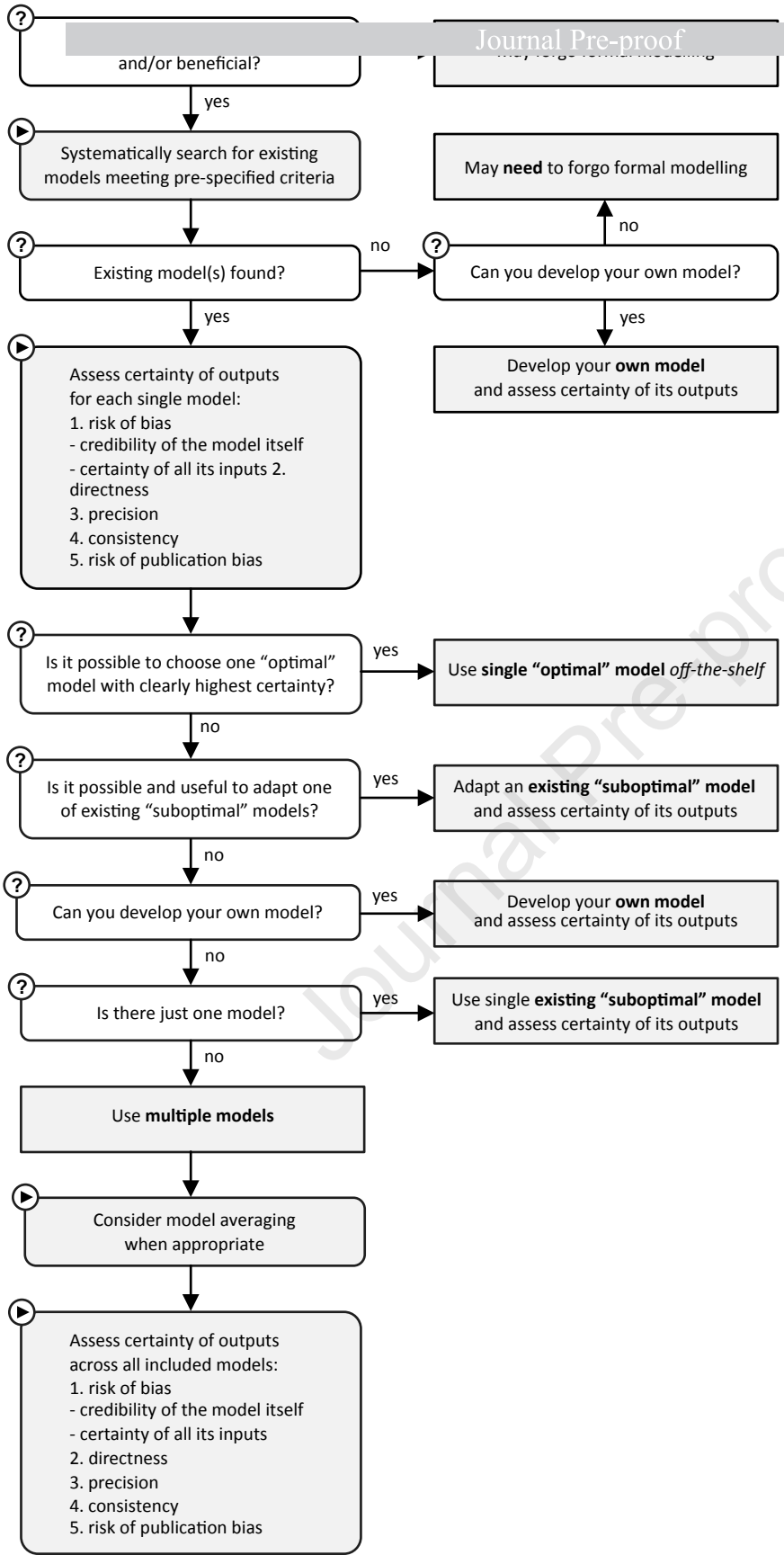
[1] Oreskes N. The role of quantitative models in science. In: Canham CD, Cole JJ, Lauenroth WK, editors. Models in ecosystem science: Princeton University Press; 2003. p. 13–31.
[2] Frigg R, Hartmann S. Models in Science. In: Zalta EN, editor. The Stanford Encyclopedia of Philosophy (Spring 2017 Edition)2017.
[3] Guyatt GH, Oxman AD, Kunz R, Vist GE, Falck-Ytter Y, Schunemann HJ, et al. What is "quality of evidence" and why is it important to clinicians? BMJ. 2008;336:995-8.
[4] Oreskes N. Evaluation (not validation) of quantitative models. Environ Health Perspect. 1998;106 Suppl 6:1453-60.
[5] Briggs AH, Weinstein MC, Fenwick EA, Karnon J, Sculpher MJ, Paltiel AD, et al. Model parameter estimation and uncertainty: a report of the ISPOR-SMDM Modeling Good Research Practices Task Force--6. Value Health. 2012;15:835-42.
[6] Caro JJ, Briggs AH, Siebert U, Kuntz KM, Force I-SMGRPT. Modeling good research practices--overview: a report of the ISPOR-SMDM Modeling Good Research Practices Task Force-1. Med Decis Making. 2012;32:667-77.
[7] Caro JJ, Eddy DM, Kan H, Kaltz C, Patel B, Eldessouki R, et al. Questionnaire to assess relevance and credibility of modeling studies for informing health care decision making: an ISPOR-AMCP-NPC Good Practice Task Force report. Value Health. 2014;17:174-82.
[8] Eddy DM, Hollingworth W, Caro JJ, Tsevat J, McDonald KM, Wong JB, et al. Model transparency and validation: a report of the ISPOR-SMDM Modeling Good Research Practices Task Force-7. Med Decis Making. 2012;32:733-43.
[9] Karnon J, Stahl J, Brennan A, Caro JJ, Mar J, Moller J. Modeling using discrete event simulation: a report of the ISPOR-SMDM Modeling Good Research Practices Task Force-4. Med Decis Making. 2012;32:701-11.
[10] Marshall DA, Burgos-Liz L, MJ IJ, Crown W, Padula WV, Wong PK, et al. Selecting a dynamic simulation modeling method for health care delivery research-part 2: report of the ISPOR Dynamic Simulation Modeling Emerging Good Practices Task Force. Value Health. 2015;18:147-60.
[11] Marshall DA, Burgos-Liz L, MJ IJ, Osgood ND, Padula WV, Higashi MK, et al. Applying dynamic simulation modeling methods in health care delivery research-the SIMULATE checklist: report of the ISPOR simulation modeling emerging good practices task force. Value Health. 2015;18:5-16.
[12] Pitman R, Fisman D, Zaric GS, Postma M, Kretzschmar M, Edmunds J, et al. Dynamic transmission modeling: a report of the ISPOR-SMDM Modeling Good Research Practices Task Force Working Group-5. Med Decis Making. 2012;32:712-21.
[13] Roberts M, Russell LB, Paltiel AD, Chambers M, McEwan P, Krahn M, et al. Conceptualizing a model: a report of the ISPOR-SMDM Modeling Good Research Practices Task Force-2. Med Decis Making. 2012;32:678-89.
[14] Siebert U, Alagoz O, Bayoumi AM, Jahn B, Owens DK, Cohen DJ, et al. State-transition modeling: a report of the ISPOR-SMDM Modeling Good Research Practices Task Force-3. Med Decis Making. 2012;32:690-700.
[15] Vemer P, van Voom GA, Ramos IC, Krabbe PF, Al MJ, Feenstra TL. Improving model validation in health technology assessment: comments on guidelines of the ISPOR-SMDM modeling good research practices task force. Value Health. 2013;16:1106-7.

[16] Weinstein MC, O'Brien B, Hornberger J, Jackson J, Johannesson M, McCabe C, et al. Principles of good practice for decision analytic modeling in health-care evaluation: report of the ISPOR Task Force on Good Research Practices--Modeling Studies. Value Health. 2003;6:9-17.
[17] Bennett C, Manuel DG. Reporting guidelines for modelling studies. BMC Med Res Methodol. 2012;12:168.
[18] Peñaloza Ramos MC, Barton P, Jowett S, Sutton AJ. A Systematic Review of Research Guidelines in Decision-Analytic Modeling. Value Health. 2015;18:512-29.
[19] Philips Z, Bojke L, Sculpher M, Claxton K, Golder S. Good practice guidelines for decision-analytic modelling in health technology assessment: a review and consolidation of quality assessment. Pharmacoeconomics. 2006;24:355-71.
[20] LaKind JS, O'Mahony C, Armstrong T, Tibaldi R, Blount BC, Naiman DQ. ExpoQual: Evaluating measured and modeled human exposure data. Environ Res. 2019;171:302-12.
[21] Husereau D, Drummond M, Petrou S, Carswell C, Moher D, Greenberg D, et al. Consolidated Health Economic Evaluation Reporting Standards (CHEERS) statement. BMJ. 2013;346:f1049.
[22] Balshem H, Helfand M, Schunemann HJ, Oxman AD, Kunz R, Brozek J, et al. GRADE guidelines: 3. Rating the quality of evidence. J Clin Epidemiol. 2011;64:401-6.
[23] Guyatt GH, Oxman AD, Kunz R, Brozek J, Alonso-Coello P, Rind D, et al. GRADE guidelines: 6. Rating the quality of evidence--imprecision. J Clin Epidemiol. 2011;64:1283-93.
[24] Guyatt GH, Oxman AD, Kunz R, Woodcock J, Brozek J, Helfand M, et al. GRADE guidelines: 8. Rating the quality of evidence--indirectness. J Clin Epidemiol. 2011;64:1303-10.
[25] Guyatt GH, Oxman AD, Kunz R, Woodcock J, Brozek J, Helfand M, et al. GRADE guidelines: 7. Rating the quality of evidence--inconsistency. J Clin Epidemiol. 2011;64:1294-302.
[26] Guyatt GH, Oxman AD, Montori V, Vist G, Kunz R, Brozek J, et al. GRADE guidelines: 5. Rating the quality of evidence--publication bias. J Clin Epidemiol. 2011;64:1277-82.
[27] Guyatt GH, Oxman AD, Sultan S, Glasziou P, Akl EA, Alonso-Coello P, et al. GRADE guidelines: 9. Rating up the quality of evidence. J Clin Epidemiol. 2011;64:1311-6.
[28] Guyatt GH, Oxman AD, Vist G, Kunz R, Brozek J, Alonso-Coello P, et al. GRADE guidelines: 4. Rating the quality of evidence--study limitations (risk of bias). J Clin Epidemiol. 2011;64:407-15.
[29] Lasserson TJ, Thomas J, Higgins JPT. Chapter 1: Starting a review. In: Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, et al., editors. Cochrane Handbook for Systematic Reviews of Interventions version 60 (updated July 2019): Cochrane; 2019.
[30] Eykhoff P. System identification: parameter and state estimation: Wiley-Interscience; 1974.
[31] Schunemann HJ, Best D, Vist G, Oxman AD, Group GW. Letters, numbers, symbols and words: how to communicate grades of evidence and recommendations. CMAJ. 2003;169:677-80.
[32] Schunemann HJ, Mustafa R, Brozek J, Santesso N, Alonso-Coello P, Guyatt G, et al. GRADE Guidelines: 16. GRADE evidence to decision frameworks for tests in clinical practice and public health. J Clin Epidemiol. 2016;76:89-98.
[33] Schunemann HJ, Oxman AD, Brozek J, Glasziou P, Jaeschke R, Vist GE, et al. Grading quality of evidence and strength of recommendations for diagnostic tests and strategies. BMJ. 2008;336:1106-10.

[34] Iorio A, Spencer FA, Falavigna M, Alba C, Lang E, Burnand B, et al. Use of GRADE for assessment of evidence about prognosis: rating confidence in estimates of event rates in broad categories of patients. BMJ. 2015;350:h870.
[35] Hooijmans CR, de Vries RBM, Ritskes-Hoitinga M, Rovers MM, Leeflang MM, IntHout J, et al. Facilitating healthcare decisions by assessing the certainty in the evidence from preclinical animal studies. PLoS One. 2018;13:e0187271.
[36] Brunetti M, Shemilt I, Pregno S, Vale L, Oxman AD, Lord J, et al. GRADE guidelines: 10. Considering resource use and rating the quality of economic evidence. J Clin Epidemiol. 2013;66:140-50.
[37] Zhang Y, Alonso-Coello P, Guyatt GH, Yepes-Nunez JJ, Akl EA, Hazlewood G, et al. GRADE Guidelines: 19. Assessing the certainty of evidence in the importance of outcomes or values and preferences-Risk of bias and indirectness. J Clin Epidemiol. 2018.
[38] Zhang Y, Coello PA, Guyatt GH, Yepes-Nunez JJ, Akl EA, Hazlewood G, et al. GRADE guidelines: 20. Assessing the certainty of evidence in the importance of outcomes or values and preferences-inconsistency, imprecision, and other domains. J Clin Epidemiol. 2018.
[39] World Health Organization. WHO guidelines for screening and treatment of precancerous lesions for cervical cancer prevention. Geneva, Switzerland: World Health Organization; 2013.
[40] Thayer KA, Schunemann HJ. Using GRADE to respond to health questions with different levels of urgency. Environ Int. 2016;92-93:585-9.
[41] Porgo TV, Norris SL, Salanti G, Johnson LF, Simpson JA, Low N, et al. The use of mathematical modeling studies for evidence synthesis and guideline development: A glossary. Res Synth Methods. 2019;10:125-33.
[42] (NICE) NIfHaCE. The reference case. Guide to the methods of technology appraisal 2013: NICE; 2013.
[43] Eyles H, Ni Mhurchu C, Nghiem N, Blakely T. Food pricing strategies, population diets, and non-communicable disease: a systematic review of simulation studies. PLoS Med. 2012;9:e1001353.
[44] Jaime Caro J, Eddy DM, Kan H, Kaltz C, Patel B, Eldessouki R, et al. Questionnaire to assess relevance and credibility of modeling studies for informing health care decision making: an ISPOR-AMCP-NPC Good Practice Task Force report. Value Health. 2014;17:174-82.
[45] (NICE) NIfHaCE. Appendix G: Methodology checklist: economic evaluations. The guidelines manual: NICE; 2012.
[46] Schultz TW, Richarz A-N, Cronin MTD. Assessing uncertainty in read-across: Questions to evaluate toxicity predictions based on knowledge gained from case studies. Computational Toxicology. 2019;9:1-11.
[47] Cronin MTD, Richarz AN, Schultz TW. Identification and description of the uncertainty, variability, bias and influence in quantitative structure-activity relationships (QSARs) for toxicity prediction. Regul Toxicol Pharmacol. 2019;106:90-104.
[48] Brazier J, Ara R, Azzabi I, Busschbach J, Chevrou-Severac H, Crawford B, et al. Identification, Review, and Use of Health State Utilities in Cost-Effectiveness Models: An ISPOR Good Practices for Outcomes Research Task Force Report. Value Health. 2019;22:267-75.
[49] Kaltenthaler E, Tappenden P, Paisley S, Squires H. NICE DSU Technical Support Document 13: Identifying and Reviewing Evidence to Inform the Conceptualisation and Population of Cost-Effectiveness Models. London2011.

[50] Paisley S. Identification of Evidence for Key Parameters in Decision-Analytic Models of Cost Effectiveness: A Description of Sources and a Recommended Minimum Search Requirement. Pharmacoeconomics. 2016;34:597-608.
[51] Guyatt G, Oxman AD, Sultan S, Brozek J, Glasziou P, Alonso-Coello P, et al. GRADE guidelines: 11. Making an overall rating of confidence in effect estimates for a single outcome and for all outcomes. J Clin Epidemiol. 2013;66:151-7.
[52] Bilcke J, Beutels P, Brisson M, Jit M. Accounting for methodological, structural, and parameter uncertainty in decision-analytic models: a practical guide. Med Decis Making. 2011;31:675-92.
[53] Saltelli A, Ratto M, Andres T, Campolongo F, Cariboni J, Gatelli D, Saisana M& Tarantola S. 2008. Global sensitivity analysis. The primer. Chichester, UK: John Wiley & Sons.
[54] Page MJ, Higgins JPT, Sterne JAC. Chapter 13: Assessing risk of bias due to missing results in a synthesis. In: Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, et al., editors. Cochrane Handbook for Systematic Reviews of Interventions version 60 (updated July 2019): Cochrane; 2019.
[55] Schünemann HJ, Lerda D, Quinn C, Follmann M, Alonso-Coello P, Rossi PG, et al. Breast Cancer Screening and Diagnosis: A Synopsis of the European Breast Guidelines. Annals of Internal Medicine. 2020;172:46-56.
[56] Eaton JW, Johnson LF, Salomon JA, Barnighausen T, Bendavid E, Bershteyn A, et al. HIV treatment as prevention: systematic comparison of mathematical models of the potential impact of antiretroviral therapy on HIV incidence in South Africa. PLoS Med. 2012;9:e1001245.
[57] Gomersall JS, Jadotte YT, Xue Y, Lockwood S, Riddle D, Preda A. Conducting systematic reviews of economic evaluations. Int J Evid Based Healthc. 2015;13:170-8.
[58] Mandelblatt JS, Stout NK, Schechter CB, van den Broek JJ, Miglioretti DL, Krapcho M, et al. Collaborative Modeling of the Benefits and Harms Associated With Different U.S. Breast Cancer Screening Strategies. Ann Intern Med. 2016;164:215-25.
[59] Davies NG, Kucharski AJ, Eggo RM, Gimma A, Edmunds WJ, Centre for the Mathematical Modelling of Infectious Diseases C-wg. Effects of non-pharmaceutical interventions on COVID-19 cases, deaths, and demand for hospital services in the UK: a modelling study. Lancet Public Health. 2020;5:e375-e85.
[60] Tibaldi R, ten Berge W, Drolet D. Dermal absorption of chemicals: estimation by IH SkinPerm. J Occup Environ Hyg. 2014;11:19-31.
[61] Young BM, Tulve NS, Egeghy PP, Driver JH, Zartarian VG, Johnston JE, et al. Comparison of four probabilistic models (CARES((R)), Calendex, ConsExpo, and SHEDS) to estimate aggregate residential exposures to pesticides. J Expo Sci Environ Epidemiol. 2012;22:522-32.
[62] United States Environmental Protection Agency. Human Exposure Modeling - Overview. In: United States Environmental Protection Agency, editor.
[63] Levin S, Dugas A, Gurses A, Kirsch T, Kelen G, Hinson J, et al. HOPSCORE: AN ELECTRONIC OUTCOMES-BASED EMERGENCY TRIAGE SYSTEM. Agency for Healthcare Research and Quality; 2018.
[64] Smith RD, Keogh-Brown MR, Barnett T, Tait J. The economy-wide impact of pandemic influenza on the UK: a computable general equilibrium modelling experiment. BMJ. 2009;339:b4571.
[65] Hultcrantz M, Rind D, Akl EA, Treweek S, Mustafa RA, Iorio A, et al. The GRADE Working Group clarifies the construct of certainty of evidence. J Clin Epidemiol. 2017;87:4-13.

869



What is new

1. General concepts determining the certainty of evidence in the GRADE approach (risk of bias, indirectness, inconsistency, imprecision, reporting bias, magnitude of an effect, dose-response relation, and the direction of residual confounding) also apply in the context of assessing the certainty of evidence from models (model outputs).
2. Detailed assessment of the certainty of evidence from models differs for the assessment of outputs from a single model compared to the assessment of outputs across multiple models.
3. We propose a framework for selecting the best available evidence from models to inform health care decisions: to develop a model de novo, to identify an existing model the outputs of which provide the highest certainty evidence, or to use outputs from multiple models.
4. We suggest that the modelling and health care decision making communities collaborate further to clarify terminology used in the context of modelling and make it consistent across the disciplines to facilitate communication.