Title

Contribution and quality of mathematical modeling evidence in World Health Organization guidelines: A systematic review

Authors

Lo, Nathan C
Andrejko, Kristin
Shukla, Poojan
et al.

Publication Date

2022-06-01

DOI

10.1016/j.epidem.2022.100570

Contribution and quality of mathematical modeling evidence in World Health Organization guidelines: a systematic review

Nathan C. Lo MD PhD[1], Kristin Andrejko BS[2], Poojan Shukla[1], Tess Baker BS[2], Veronica Ivey Sawin MS[3], Susan L. Norris MD MPH[3*], and Joseph A. Lewnard PhD[2,4,5]

[1] Division of HIV, Infectious Diseases, and Global Medicine, University of California, San Francisco, San Francisco, California, USA
[2]Division of Epidemiology, University of California, Berkeley, School of Public Health, Berkeley, California, USA
[3]Department of Quality of Norms and Standards, Science Division, World Health Organization, Geneva, Switzerland
[4]Division of Infectious Diseases and Vaccinology, University of California, Berkeley, School of Public Health, Berkeley, California, USA
[5]Center for Computational Biology, College of Engineering, University of California, Berkeley, Berkeley, California, USA
*Affiliation at the time this work was performed

Abstract Word Count: 345

Main Text Word Count: 3,666

Figures: 3

Tables: 3

Keywords: mathematical modeling, guidelines, World Health Organization, public health, health policy, evidence-based medicine

Correspondence:
Nathan C. Lo, MD PhD
Division of HIV, Infectious Diseases, and Global Medicine
University of California, San Francisco
San Francisco, CA, USA
Nathan.Lo@ucsf.edu

ABSTRACT

Mathematical modeling studies are frequently conducted to guide policy in global health. However, the contribution of mathematical modeling studies to World Health Organization (WHO) guideline recommendations, and the quality of evidence contributed by these studies remains unknown. We conducted a systematic review of the WHO Guidelines Review Committee database to identify guideline recommendations that included evidence from mathematical modeling studies since inception of the Guidelines Review Committee on 1 December, 2007. We included WHO guideline recommendations citing a mathematical modeling study in the primary evidence base. We defined a mathematical model as a framework that predicted epidemiologic, health or economic impact of an intervention or decision in the clinical or public health context. The primary outcome was inclusion of evidence from mathematical modeling studies in a guideline recommendation. We evaluated each unique modeling study across multiple domains of quality. Between 1 December 2007 and 1 April 2019, the WHO Guidelines Review Committee approved 154 guidelines providing 1,619 guideline recommendations. Mathematical modeling studies informed 46 WHO guidelines (29.9%) and 101 unique guideline recommendations (6.2%). Modeling evidence addressed topics related to infectious diseases in 38 guidelines (82.6%) and 81 recommendations (80.2%), most commonly for HIV and tuberculosis. Evidence from modeling studies was assessed in the GRADE evidence profile for 12 recommendations (12.9%) and GRADE evidence-to-decision framework for 45 recommendations (44.6%). Modeling-informed recommendations were more likely than other recommendations within the same guidelines to be issued with a "conditional" rather than "strong" strength of recommendation (53.5% versus 37.8%), and the evidence underlying

modeling-informed recommendations was more likely to be assessed as very low quality (41.6% versus 24.1%). Upon review of individual modeling studies, we estimated that 33.8% of models performed a calibration, 29.4% of models performed a validation of results, and 20.6% of models reported a change in the study conclusion in the sensitivity analysis. While policy recommendations in WHO guidelines are informed by evidence from modeling studies, the validity of modeling studies included in guidelines development is heterogeneous. Quality assessment is needed to support the evaluation and incorporation of evidence from mathematical modeling studies in guidelines development.

Introduction

Mathematical modeling studies are frequently undertaken with the aim of informing clinical and public health policy.[1,2] Over the past decade, these studies have become increasingly common and relied upon in decision-making in global public health, especially during the COVID-19 pandemic.[1,2] While mathematical modeling encompasses a diverse set of methodologies, these studies can often be defined by their application of a mathematical framework to predict epidemiologic, health, or economic impacts of an intervention or decision in a clinical or public health context, usually when direct observation or measurement of such impacts is infeasible. In these respects, mathematical modeling studies are distinct from routine statistical analyses of observed data that may use statistical models to facilitate hypothesis testing or estimation.

While the potential uses of modeling for evidence generation are vast and varied, results of modeling studies may be of particular relevance to policymakers when conventional epidemiologic evidence (e.g., contributed by direct observations from randomized or non-randomized studies) is unavailable. This may occur when the design of such studies is impractical (e.g., due to the duration of follow up or sample size required) or unethical (e.g., in the case of withholding an efficacious intervention from study participants), or when the study question is not well suited for traditional epidemiologic assessment (e.g., balancing resources and benefits). Some examples of mathematical modeling studies include: i) comparison of new diagnostic tools for a screening program; ii) an outbreak prediction to guide a public health

response[3]; iii) estimation of the impact of an intervention in a specific population, over a long-term period, and potential spill-over benefit (e.g., screening for preventive care, vaccine introduction)[4]; or iv) evaluation of the economic impact or cost-effectiveness of a decision (e.g., new medication for chronic diseases, cost-effectiveness of a new program).[5]

The World Health Organization (WHO) has the mandate to establish standards in global health through the production of normative, technical guidance based on the best available evidence.[6] WHO publishes guidelines across a broad range of clinical, public health, and policy topics, providing specific recommendations for clinical care and for public health strategies and programmes.[7,8] These guidelines are highly influential in changing clinical care and public health. Some examples include WHO guidelines that changed the CD4 treatment threshold for HIV in 2009 and 2013, greatly expanding access to antiretroviral therapy in low- and middle-income countries;[9] 2006 guidelines that initiated large-scale mass treatment programs against six neglected tropical diseases; [10]1998 guidelines that addressed global tobacco control;[11] and informing countries' adoption of vaccines through the Expanded Programme on Immunization.[12]

WHO guideline recommendations are based on a systematic review of the evidence on benefits and harms for each topic, and on the quality (or certainty) of evidence assessed using the Grading of Recommendations, Assessment, Development and Evaluations (GRADE) system. The GRADE assessment of high, moderate, low, or very low quality evidence is usually presented in an "Evidence Profile" alongside each recommendation and includes five key domains: study limitations, inconsistency, indirectness, imprecision, and publication bias. Recommendations are assigned a strength of strong or conditional (weak) based on a structured

evidence-to-decision framework that includes a variety of decision factors.[7,13,14] However, modeling methods and evidence are not formally addressed by the GRADE system and there are no formal reporting and assessment standards to guide the incorporation of mathematical modeling.[15]

To understand the contribution of mathematical modeling studies to WHO guidelines, and the quality of included studies, we performed a systematic review of WHO guidelines which have been reviewed and approved by the Guidelines Review Committee (GRC), the quality assurance body for WHO guidelines.

Methods

Search strategy and selection criteria

In this systematic review, we sought to identify all WHO GRC-approved guideline recommendations that included evidence from mathematical modeling studies. This study follows the Preferred Reporting Items for Systematic reviews and Meta-Analysis (PRISMA) guidelines.[16] The systematic review protocol, including the analytical plan, is available in Supplemental materials.

We obtained the WHO GRC database from the GRC Secretariat, and searched it to identify a complete list of all GRC-approved documents from inception of the GRC (1 December, 2007) to 1 April, 2019.[7,8] We excluded documents characterized as information products including "how to" or "information documents", as defined in the *WHO Handbook on Guideline Development*.[3]

We aimed to identify all WHO GRC-approved guideline recommendations that included citation to a mathematical modeling study in the main evidence base, including the GRADE Evidence Profile/Table, GRADE Evidence-to-Decision Framework, summary of evidence section, or supplemental section (defined as not meeting inclusion in a previous category).[7,14,17] We defined mathematical modeling studies as those using a mathematical framework to predict the epidemiologic, health, or economic impact of an intervention or decision in the clinical or public health context, in the absence of direct observations of such impacts from either randomized or non-randomized assessments. We used a free text search term of "model*" to capture all potential mentions of mathematical modeling evidence. This single search term was sensitive in capturing a broad range of mathematical modeling studies, although we piloted additional search terms, including "simulat*" (for simulation) and "cost-effect*" (for cost-effectiveness), in the piloted subset of guidelines. WHO guidelines do not have keywords or controlled vocabulary (i.e., MeSH terms).

We excluded guideline recommendations in the following situations: (i) the guideline was not submitted to the GRC in English; (ii) the recommendation did not include an assessment of the quality of the evidence or indicate the strength of recommendation; (iii) the recommendation incorporated only evidence from routine statistical models (including generalized linear models, generalized estimating equations, random/fixed effects models for meta-analysis, and conventional statistical tests) or qualitative conceptual models (e.g., logic model, implementation science model) without additional exploratory quantitative frameworks; (iv) modeling evidence was not applied in the main evidence base (for instance, guidelines citing modeling studies in the Background section); and (v) the recommendation had been re-issued in successive guideline

iterations using a duplicate reference for the same modeling studies. The inclusion and exclusion criteria were applied at the level of WHO guideline recommendations (i.e., a single WHO guideline may have multiple recommendations, however only a subset of the recommendations may meet inclusion for the study). Two reviewers independently performed full-text screening. Any disagreement was resolved through discussion between the two reviewers, with involvement of a third reviewer as necessary. In the review of individual modeling studies included in the systematic review, only studies with available published reports or studies were evaluated.

Data extraction and outcomes of WHO guidelines

The first reviewer extracted data from each guideline recommendation into a tabular format, and a second reviewer independently verified the data extracted by the first reviewer. The data that were extracted included: (i) where modeling was presented in the guideline (in the GRADE Evidence Profile/Table, GRADE Evidence-to-Decision Framework, summary of evidence section, or supplemental section); (ii) topic area of the guideline; (iii) type of question (e.g., economic, intervention effect); (iv) quality of evidence for the recommendation (very low, low, moderate, or high); (v) strength of recommendation (conditional/weak, strong); (vi) GRADE evidence-to-decision criteria (i.e., priority of problem, test accuracy, benefits and harms, values and preferences, acceptability, resource implications, equity, and feasibility) when applicable; (vii) source of the model (previously published or newly commissioned); (viii) comparison of results across independent models (where applicable); (ix) availability of non-modeling evidence. We extracted the reporting of any key factors affecting quality of the modeling evidence, including model assumptions and limitations, sensitivity analyses, and model validation. If multiple distinct models informed a single guideline recommendation, we assessed

whether the recommendation included text addressing assumptions and limitations, sensitivity analyses, and validation for each of the included models; we calculated an average for the reporting of each of these items in each guideline based on the proportion of models for which assessments were undertaken. Among guidelines that included at least one recommendation informed by modeling evidence, the quality of evidence and strength of recommendation for the remaining guideline recommendations that were not informed by modeling was also extracted. Across the guidelines including modeling, the topic area of each guideline and the total count of recommendations informed or not informed by modeling was extracted.

For each guideline recommendation, the primary outcome was the inclusion of modeling evidence; we assessed the proportion of recommendations including modeling evidence. For those recommendations that included modeling evidence, we further assessed where results of modeling studies were presented within guidelines, and how the incorporation of evidence from modeling studies was described.

Data extraction and outcomes of mathematical modeling studies

We reviewed each mathematical modeling study applied in a WHO guideline recommendation using a standardized survey instrument to evaluate across domains that may predict quality of evidence. We developed the survey to evaluate characteristics of each modeling study based on a review of published literature addressing quality and reporting of modeling studies including input from the International Society for Pharmacoeconomics and Outcomes Research (ISPOR).[18-21] The survey addressed the following characteristics: model structure and assumptions, calibration, influential model inputs, robustness of sensitivity and uncertainty analyses,

validation, face validity, generalizability, inclusion of pre-analysis plan, and conflicts of interest. The list of survey questions and criteria for evaluation is available in the Supplemental materials**.** Two independent reviewers extracted data from each mathematical modeling study into a tabular format with input from the team; in cases of disagreement, a third reviewer resolved the decision.

Statistical analysis

We reported descriptive totals and proportions of recommendations meeting the criteria described above. For outcomes related to modeling-informed recommendations alone, we used this total as the denominator in the reported proportions; to make this clear we present numerators and denominators for these estimates. We made inferential comparisons of evidence quality and recommendation strength for recommendations that were, and were not, informed by models; we made these comparisons within the subset of guidelines that included at least one model-informed recommendation in order to control for potential differences in assessment across subject areas and guidelines committees. We computed a χ^2 test statistic to assess differences in proportions for quality of evidence and strength of recommendation across recommendations informed and not informed by modeling; we used the Fisher exact test for cell sizes with fewer than five recommendations. Hypothesis testing was conducted at a two-sided significance level of $p<0.05$.

Reviewers recorded data in a Microsoft Excel spreadsheet (version 16.30, Microsoft, Redmond, WA); statistical analysis and data visualization were done with R 3.2.3 (R Foundation for

Statistical Computing; Vienna, Austria). The authors support the importance of data sharing and transparency in research; data and analysis code are fully accessible in an online repository.[22]

Results

Summary of guidelines and recommendations

Between 1 December 2007 and 1 April 2019, the WHO GRC approved 250 guideline documents, of which 154 were eligible guidelines that fulfilled inclusion criteria for full-text review; 57 of 96 excluded documents did not present clear evidence-based recommendations, 35 were informational, and 4 were not published (**Figure 1**). The 154 included guidelines presented 1,619 unique recommendations. Evidence from mathematical modeling studies informed 101 unique guideline recommendations (representing 6.2% of all 1,619 guideline recommendations) issued in 46 WHO GRC-approved guidelines (representing 29.9% of all 154 included guidelines). These 46 included GRC-approved guidelines supported 461 total recommendations, in total. We present a full description of the exclusion process in the Supplemental materials.

We identified a total of 113 unique mathematical models cited in the 101 guideline recommendations, with a subset of models cited across multiple guidelines and recommendations. An average of 1.6 (median: 1, range: 1-9) unique modeling studies were cited by each recommendation that incorporated modeling evidence; 28 (27.8% of 101) modeling-informed recommendations cited more than one mathematical model. We identified the number of guidelines and guideline recommendations informed by mathematical modeling each year in

Figure 2; no clear trend over time was evident in the proportion of guidelines or guideline recommendations that incorporated modeling evidence.

Use of modeling

The majority of guideline recommendations informed by modeling evidence addressed topics in infectious diseases (81/101; 80.2%), with the largest number addressing tuberculosis (27/101; 24.5%) and HIV (24/101; 23.8%) (**Figure 3**). Of the total recommendations (N= 588) published in HIV and TB guidelines between 2008 and 2019, modelling evidence was utilized to support decisions in 6.0% (24/403) and 14.6% (27/185) of recommendations, respectively. The remaining guideline recommendations informed by modeling (N=20) addressed topics in primary healthcare services, cancer screening and treatment, management of diabetes, nutrition, mental health and substance abuse, rehabilitation services for persons with disabilities, health products, and environmental risk management. The majority of guideline recommendations including modeling evidence concerned recommendations about intervention effects (68/101; 67.3%) or diagnostic testing (17/101; 16.8%).

In total, 12 guideline recommendations (11.9% of 101 modeling-informed recommendations) from 5 guidelines assessed modeling evidence in the GRADE Evidence Profile (**Table 1**). In 45 guideline recommendations (44.6% of 101) from 27 guidelines, modeling studies were included in the GRADE Evidence-to-Decision Framework, with the most common decision criteria being resource implications (80.0% of 45) and benefits and harms (24.4% of 45). Most modeling-informed guideline recommendations relied on additional observational or experimental

evidence, in addition to the modeling work, to support the guideline recommendation; only four guideline recommendations from two guidelines were based on modeling evidence alone, and all evidence quality was rated very low or low. For each of the five guideline recommendations, modeling evidence was reviewed in the GRADE Evidence Profile (**Table S2**).

Overall, the distribution of quality of evidence ratings differed for recommendations from the same guidelines that included, or did not include, mathematical modeling studies ($\chi^2_{df=3}=15.72$; $p<0.01$). The majority of modeling-informed guideline recommendations were judged to have very low (42/101; 41.6%) or low quality of evidence (36/101 35.6%); only a minority were judged to have moderate quality (16/101; 15.8%) or high quality (7/101; 6.9%) evidence (**Table 2**). In comparison, recommendations not informed by modeling evidence were less likely to be assessed to have very low-quality evidence (90/373 [24.1%]; $p<0.01$), and were approximately twice as likely to have moderate- or high-quality evidence (148/373 [39.7%] versus 23/101 [22.8%]; $p<0.01$). Modeling-informed recommendations not related to infectious diseases were more likely to be very low or low quality (18/20; 90.0%) than modeling-informed recommendation related to infectious disease (60/81; 74.1%).

Recommendations informed by modeling evidence were also more likely to be issued with a "conditional" strength of recommendation (also known as "weak" strength) than recommendations from the same guidelines not informed by modeling evidence (54/101 [53.5%] versus 141/373 [37.8%]; $p<0.01$). Among recommendations that were issued with a "strong" designation, differences in the distribution of evidence ratings were not evident between

recommendations informed by modeling studies or not informed by modeling studies ($\chi^2_{df=3}$=4.45; p=0.21). Modeling-informed recommendations related to infectious diseases had similar proportion of conditional strength of recommendation (43/81; 53.1%) when compared to non-infectious disease modeling-informed recommendations (11/20; 55.0%).

We also evaluated the reporting of mathematical modeling evidence within guideline recommendation documents. Model assumptions or limitations were reported in 93 recommendations (93/101; 92.1%) and sensitivity analysis or robustness of model was cited in 68 (67.3% of 101) of recommendations. Only 2 recommendations (2.0% of 101) reported attempts to validate the modeling study results. A similar proportion of recommendations using mathematical modeling evidence relied upon studies that were commissioned (46/101; 45.5%) and existing, published models (48/101; 47.5%); 7 relied on both types of studies (7/101; 6.9%). Guideline recommendations with models that were commissioned were more likely to cite assumptions or limitations (45/46 [97.8%] versus 41/48 [85.4%]; p=0.07), and to address the results of sensitivity analyses (43/46 [93.5%] versus 20/48 [41.7%]; p<0.01).

Quality of individual modeling studies

We performed an in-depth evaluation of a total of 68 unique mathematical modeling studies, after exclusion of 50 (42.4% of all modeling studies forming the evidence base for WHO guideline recommendations) that were not available in a published or accessible form (Table 3). The majority of modeling studies were judged to have reasonable model structure and assumptions (67/68; 98.5%) with sufficient description (66/68; 97.1%), although fewer provided

a full mathematical description of the model (56/68; 82.4%). Only 33.8% (23/68) of models performed calibration, and only 25% (~74% of those performing calibration) demonstrated consistency with observed data. The conclusion was sensitive to 1-2 key model inputs in approximately 20% (14/68) of models. We identified that only 29.4% (20/68) of models presented a validation of their results, although all were judged to have face validity. Only two modeling studies (2.9%) included a pre-analysis plan. A conflict of interest was present in 10.3% (7/68) of studies.

Discussion

In this systematic review, we found that mathematical modeling studies have been used to inform almost a third of all WHO guidelines and 6% of all WHO guideline recommendations. The majority of modeling-informed recommendations addressed topics in infectious diseases, although other areas of global health were also represented. The evidence underlying modeling-informed recommendations was assessed as lower quality than the evidence underlying other recommendations. We identified that only a third of models included in guideline evidence performed a calibration and 30% of models performed a validation of results. Over 40% of all modeling studies were found to not have a published or publicly available report. These uses of modeling evidence in WHO guidelines underscore the need for a framework to evaluate the quality of evidence contributed by mathematical models. This study highlights key areas of quality improvement for modeling studies included in WHO guidelines (e.g., publication of modeling study, calibration and validation of model) and provides a survey that could be readily

adapted for systemic evaluation of modeling studies for guideline committees and other non-modeling audiences (e.g., policymakers).

GRADE provides a process for evaluating quality of evidence in WHO guidelines,[13,14] but does not currently have a framework to evaluate mathematical modeling studies.[2,13-15] Despite this, of the 47 guidelines that used modeling evidence, 5 evaluated models in the GRADE Evidence Profile using the five factors; another 27 guidelines used modeling evidence in the GRADE Evidence-to-Decision framework. However, the majority of mathematical modeling evidence did not apply either GRADE approach, which likely reduces transparency and may contribute to heterogeneity in whether and how modeling studies have been assessed and incorporated into WHO guidelines. In the few guideline recommendations that did apply GRADE to modeling studies, none reported the three upgrading dimensions (i.e., large effect, dose response, opposing bias and confounding). The large variation in reporting of key metrics (calibration, validation) and difficulty to evaluate bias is supported by a recent analysis of modeling studies applied specifically to COVID-19 prediction.[23]

Modeling evidence was found to be applied to guideline recommendations across all levels of quality of evidence and strength of recommendation. However, modeling-informed recommendations were based on evidence that was assessed as lower quality than other non-modeling-informed recommendations issued in the same guidelines, and were more likely to be conditional (versus strong), with modestly higher quality of evidence for infectious rather than non-infectious disease topics. This finding may suggest that guideline development committees are not assured of the quality of evidence presented in modeling studies, or lack a robust method

to adequately assess model quality. It is also possible that modeling evidence is sought preferentially in situations where only low- or very low-quality of evidence is available from other sources. The evaluation of quality is likely influenced by a guideline committee's confidence in the study design of mathematical model as well as limitations specific to each model.

While our analysis focused on contribution and reporting of modeling evidence in WHO guidelines, the appraisal of evidence from mathematical modeling studies is broadly applicable to both research and policy decision-making. We adapted published recommendations across multiple sources into a survey to evaluate individual modeling studies, which could be readily used to evaluate the quality of an individual mathematical modeling study; these recommendations provide both nontechnical and technical documentation regarding model parameters, structure, and development processes.[24] However, there remains limited guidance on how the quality of the body of evidence should be assessed beyond a single modeling study, and lack of systematic framework to translate quality of evidence to inform a guideline recommendation. Recent consultations at WHO have highlighted these issues and WHO has initiated the process of soliciting guidance.[15,25] The WHO Immunization and Vaccine related Implementation Research Advisory Committee (IVIR-AC) has developed some guidance on evaluation of quality of mathematical modeling as evidence, including the use of multi-model comparison studies.[26-28]

The findings of this study should be interpreted within the context of the limitations of the data and study design. We extracted information provided within the guidelines document, although

guidelines committee members may have discussed or considered modeling evidence that was not captured in the final document (e.g., assumptions and limitations, sensitivity analysis, and validation). While our search strategy was designed to be broad, it is possible the review missed uses of modeling evidence not associated with the term "model." However, alternative search terms did not yield additional results when piloted in a sample of guidelines documents. While we found that a number of WHO guidelines incorporated evidence from mathematical modeling, the relative importance of this evidence in driving recommendations was unable to be ascertained based on text review of the guidelines document. While guideline reporting of model assumptions, sensitivity analyses, and validation may indicate attention of guidelines committee members to model credibility, we were unable to ascertain how these considerations affected the decision to issue recommendations and ratings of evidence quality. Of these features, external validation may be an important predictor of accuracy, but few modeling studies informing WHO guidelines addressed external validation.[24,29,30] We applied a survey to evaluate quality of individual model studies following pre-defined criteria for each survey question, although in some cases there was uncertainty in classification. A common concern in evaluating modeling studies is a lack of transparency in complex modeling decisions and assumptions. Future efforts should determine standard of practice for methodologic explanation, data and code sharing, and transparency.

Mathematical models are frequently used to inform WHO guideline recommendations for clinical and public health practices, although evaluation and reporting of this evidence is highly varied. This study suggests that a formal process is needed to evaluate the quality of

mathematical modeling evidence, define reporting standards, and develop a framework to guide decisions on when and how modeling evidence can be incorporated into WHO guidelines.

References

1. Basu S, Andrews J. Complexity in mathematical models of public health policies: a guide for consumers of models. *PLoS Med* 2013; **10**(10): e1001540.

2. Porgo TV, Norris SL, Salanti G, et al. The use of mathematical modeling studies for evidence synthesis and guideline development: A glossary. *Res Synth Methods* 2019; **10**(1): 125-33.

3. Lewnard JA, Antillon M, Gonsalves G, Miller AM, Ko AI, Pitzer VE. Strategies to Prevent Cholera Introduction during International Personnel Deployments: A Computational Modeling Analysis Based on the 2010 Haiti Outbreak. *PLoS Med* 2016; **13**(1): e1001947.

4. Penny MA, Verity R, Bever CA, et al. Public health impact and cost-effectiveness of the RTS,S/AS01 malaria vaccine: a systematic comparison of predictions from four mathematical models. *Lancet* 2016; **387**(10016): 367-75.

5. Basu S, Wagner RG, Sewpaul R, Reddy P, Davies J. Implications of scaling up cardiovascular disease treatment in South Africa: a microsimulation and cost-effectiveness analysis. *Lancet Glob Health* 2019; **7**(2): e270-e80.

6. The World Health Organization: working for better health for everyone, everywhere. World Health Organization, 2018.

7. WHO Handbook for Guideline Development: World Health Organization 2014.

8. Norris SL, Ford N. Improving the quality of WHO guidelines over the last decade: progress and challenges. *Lancet Glob Health* 2017; **5**(9): e855-e6.

9. Richardson ET, Grant PM, Zolopa AR. Evolution of HIV treatment guidelines in high- and low-income countries: converging recommendations. *Antiviral Res* 2014; **103**: 88-93.

10. WHO. Preventive chemotherapy in human helminthiasis. Coordinated use of anthelminthic drugs in control interventions: a manual for health professionals and programme managers: Geneva: World Health Organization, 2006.

11. WHO. Guidelines for controlling and monitoring the tobacco epidemic: Geneva: World Health Organization, 1998.

12. Keja K, Chan C, Hayden G, Henderson RH. Expanded programme on immunization. *World Health Stat Q* 1988; **41**(2): 59-63.

13. Atkins D, Best D, Briss PA, et al. Grading quality of evidence and strength of recommendations. *BMJ* 2004; **328**(7454): 1490.

14. Guyatt GH, Oxman AD, Vist GE, et al. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. *BMJ* 2008; **336**(7650): 924-6.

15. Egger M, Johnson L, Althaus C, et al. Developing WHO guidelines: Time to formally include evidence from mathematical modelling studies. *F1000Res* 2017; **6**: 1584.

16. Moher D, Liberati A, Tetzlaff J, Altman DG, Group P. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *BMJ* 2009; **339**: b2535.

17. Andrews JC, Schunemann HJ, Oxman AD, et al. GRADE guidelines: 15. Going from evidence to recommendation-determinants of a recommendation's direction and strength. *J Clin Epidemiol* 2013; **66**(7): 726-35.

18. Bennett C, Manuel DG. Reporting guidelines for modelling studies. *BMC Med Res Methodol* 2012; **12**: 168.

19. 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. *Value Health* 2012; **15**(6): 796-803.

20. 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**(4): 355-71.

21. Ramos MC, Barton P, Jowett S, Sutton AJ. A Systematic Review of Research Guidelines in Decision-Analytic Modeling. *Value Health* 2015; **18**(4): 512-29.

22. Andrejko K. Publication Data and Code. https://github.com/kristinandrejko/WHO-Mathematical-Modeling. 2020.

23. Wynants L, Van Calster B, Collins GS, et al. Prediction models for diagnosis and prognosis of covid-19 infection: systematic review and critical appraisal. *BMJ* 2020; **369**: m1328.

24. Eddy DM, Hollingworth W, Caro JJ, Tsevat J, McDonald KM, Wong JB. Model Transparency and Validation: A Report of the ISPOR-SMDM Modeling Good Research Practices Task Force-7. *Value in Health*; **15**(6): 843-50.

25. Consultation on the Development of Guidance on How to Incorporate the Results of Modelling into WHO Guidelines: World Helath Organization, 2016.

26. den Boon S, Jit M, Brisson M, et al. Guidelines for multi-model comparisons of the impact of infectious disease interventions. *BMC Med* 2019; **17**(1): 163.

27. Drolet M, Benard E, Jit M, Hutubessy R, Brisson M. Model Comparisons of the Effectiveness and Cost-Effectiveness of Vaccination: A Systematic Review of the Literature. *Value Health* 2018; **21**(10): 1250-8.

28. Report on the Immunization and Vaccine related Implementation Research (IVIR) Advisory Committee Meeting. *WHO* 2018.

29. Eddy DM. Accuracy versus Transparency in Pharmacoeconomic Modelling. *PharmacoEconomics* 2006; **24**(9): 837-44.

30. Jaime Caro J, Eddy DM, Kan H, 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**(2): 174-82.

Tables and Figures

Table 1: Location of mathematical modeling evidence in WHO guideline recommendations

Location of modeling evidence	Guidelines with mathematical models	Model-based recommendations
	n	n (%)
GRADE Evidence Profile/Table	5	12 (11.9)
GRADE Evidence-to-Decision Framework	27	45 (44.6)
Summary of evidence section	13	23 (22.8)
Supplemental section	8	21 (20.8)

This table summarizes where modeling studies were identified. The supplemental section location for modeling evidence included presentation under topic-specific sub-headings that followed the recommendation. A total of 46 guidelines and 101 unique recommendations were informed by mathematical modeling evidence; the column total may not equal these values because modeling evidence could be cited in multiple sections for each guideline and recommendation. Note that in rare examples the evidence-to-decision framework was not the GRADE framework.

GRADE; Grading of Recommendations Assessment, Development and Evaluations

Table 2: Quality of evidence and strength of recommendations in WHO guidelines that incorporated evidence from mathematical modeling studies.

Strength of Recommendation										
Non-model-informed recommendations				Model-informed recommendations						
					Non-infectious Disease			Infectious Disease		
Quality of Evidence	Conditional (n =141)	Strong (n =232)	Overall (n=373)	**Quality of Evidence**	Conditional (n=11)	Strong (n=9)		Conditional (n = 43)	Strong (n = 38)	Overall (n = 101)
Very Low (n=90)	54 (38.3%)	36 (15.5%)	90 (24.1%)	Very Low (n=42)	9 (81.8%)	6 (66.7%)	HIV/TB	12 (27.9%)	6 (15.8%)	42 (41.6%)
							Other	8 (18.6%)	1 (2.6%)	
Low (n=135)	61 (43.3%)	74 (31.9%)	135 (36.2%)	Low (n=36)	2 (18.2%)	1 (11.1%)	HIV/TB	9 (20.9%)	9 (23.7%)	36 (35.6%)
							Other	10 (23.3%)	5 (13.2%)	
Moderate (n=117)	25 (17.7%)	92 (39.7%)	117 (31.4%)	Moderate (n=16)	0 (0%)	2 (22.2%)	HIV/TB	2 (4.7%)	8 (21.1%)	16 (15.8%)
							Other	0 (0%)	4 (10.5%)	
High (n=31)	1 (0.7%)	30 (12.9%)	31 (8.3%)	High (n=7)	0 (0%)	0 (0%)	HIV/TB	2 (4.7%)	3 (7.9%)	7 (6.9%)
							Other	0 (0%)	2 (5.3%)	
Overall	141 (37.8%)	232 (62.2%)		Overall	11 (10.9%)	9 (8.9%)	HIV/TB	25 (24.7%)	26 (25.7%)	
							Other	18 (17.8%)	12 (11.9%)	

"Quality of evidence" (also referred to as "certainty of evidence") refers to the assessments of a body of evidence using the GRADE system. We computed a χ^2 test statistic to assess differences in proportions in quality of evidence and strength of recommendation across recommendations informed and not informed by modeling; we used the Fisher exact test for cell sizes with fewer than five recommendations.

Hypothesis testing was conducted at a two-sided significance level to compare the proportion of recommendations rated very low (24.1% vs. 41.6%, $p<0.001$), low (26.2% vs. 25.6%, $p=1$), moderate (30.9% vs. 15.8%, $p<0.01$), or high (8.3% vs. 6.9%, $p=0.8$) in non-modeling informed vs. modeling informed recommendations. The proportion of total recommendations rated conditional was significantly lower in non-modeling recommendations than modeling recommendation (37.8% vs. 53.5%, $p<0.01$). The proportion of total recommendations rated as strong was significantly higher in non-model based recommendations than model-based recommendations (62.2% vs. 46.5%, $p<0.01$). The overall distribution of quality of evidence significantly varied between non-modeling and modeling-based recommendations ($p<0.01$)

Table 3: Evaluation of quality of individual mathematical modeling studies included in WHO guideline recommendations.

Category	Question	Yes (n = 68) (%)	No (n = 68) (%)
Model Structure and Assumptions	Is the structure of the model and key assumptions reasonable?	67 (98.5)	1 (1.5)
Model Structure and Assumptions	Is there a sufficient description of the model structure, assumptions, and limitations that support key modelling decisions?	66 (97.1)	2 (2.9)
Model Structure and Assumptions	Do the authors provide a complete mathematical description of their model?	56 (82.4)	12 (17.6)
Model Structure and Assumptions	Is there a formal process to compare alternative model structures or assumptions to inform the final model?	7 (10.3)	61 (89.7)
Calibration	Was model calibration or parameter fitting conducted?	23 (33.8)	45 (66.2)
Calibration	Is the calibrated model result broadly consistent with observed data used in calibration?	17 (25)	51 (75)
Influential Fixed Model Inputs	Are the most influential inputs in the analysis (e.g. key effect size of intervention) tested in the sensitivity analysis?	55 (80.9)	13 (19.1)
Robustness of Sensitivity Analysis	Was a one-way sensitivity analysis performed with a range of values that is reasonable?	58 (85.3)	10 (14.7)
Robustness of Sensitivity Analysis	Is the study conclusions robust in the one-way sensitivity analysis in the eyes of the reviewer?	54 (79.4)	14 (20.6)
Robustness of Uncertainty Analysis	Was a multivariate uncertainty analysis conducted in which multiple parameters are simultaneously varied by choosing values from a distribution?	39 (57.4)	29 (42.6)
Robustness of Uncertainty Analysis	Does the uncertainty analysis produce a robust 95% uncertainty interval that is broadly consistent with the study conclusion?	36 (52.9)	32 (47.1)
Face Validity	Is the model conclusion broadly agreeable with expert intuition?	68 (100)	0 (0)
External/Internal Validation	Was some form of validation with the model prediction, either external or internal, conducted?	20 (29.4)	48 (70.6)
External/Internal Validation	In the validation, was the model prediction consistent with the observed data and conducted with meaningful rigor?	15 (22.1)	53 (77.9)

Generalizability	Are the results and the model inputs relevant to the policy topic at hand?	64 (94.1)	4 (5.9)
Inclusion of a Pre-Specification Plan	Did the study include a pre-specification plan which pre-defines key aspects of model structure and fitted parameters?	2 (2.9)	66 (97.1)
Funder Conflict of Interest	Does the study disclose any secondary interests which call into question overall study conclusions?	7 (10.3)	61 (89.7)

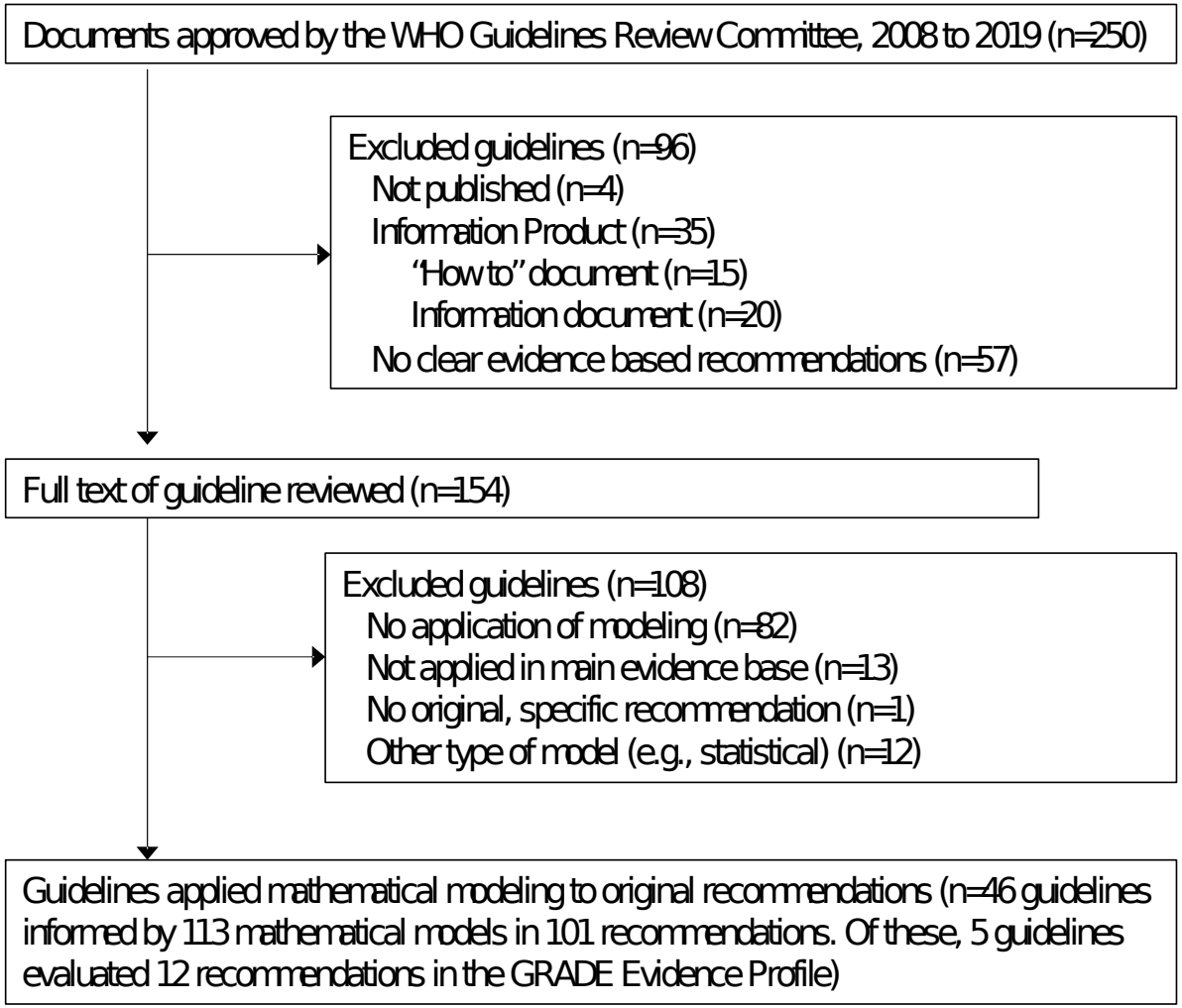

Figure 1: PRISMA flowchart of WHO guideline and recommendation selection.

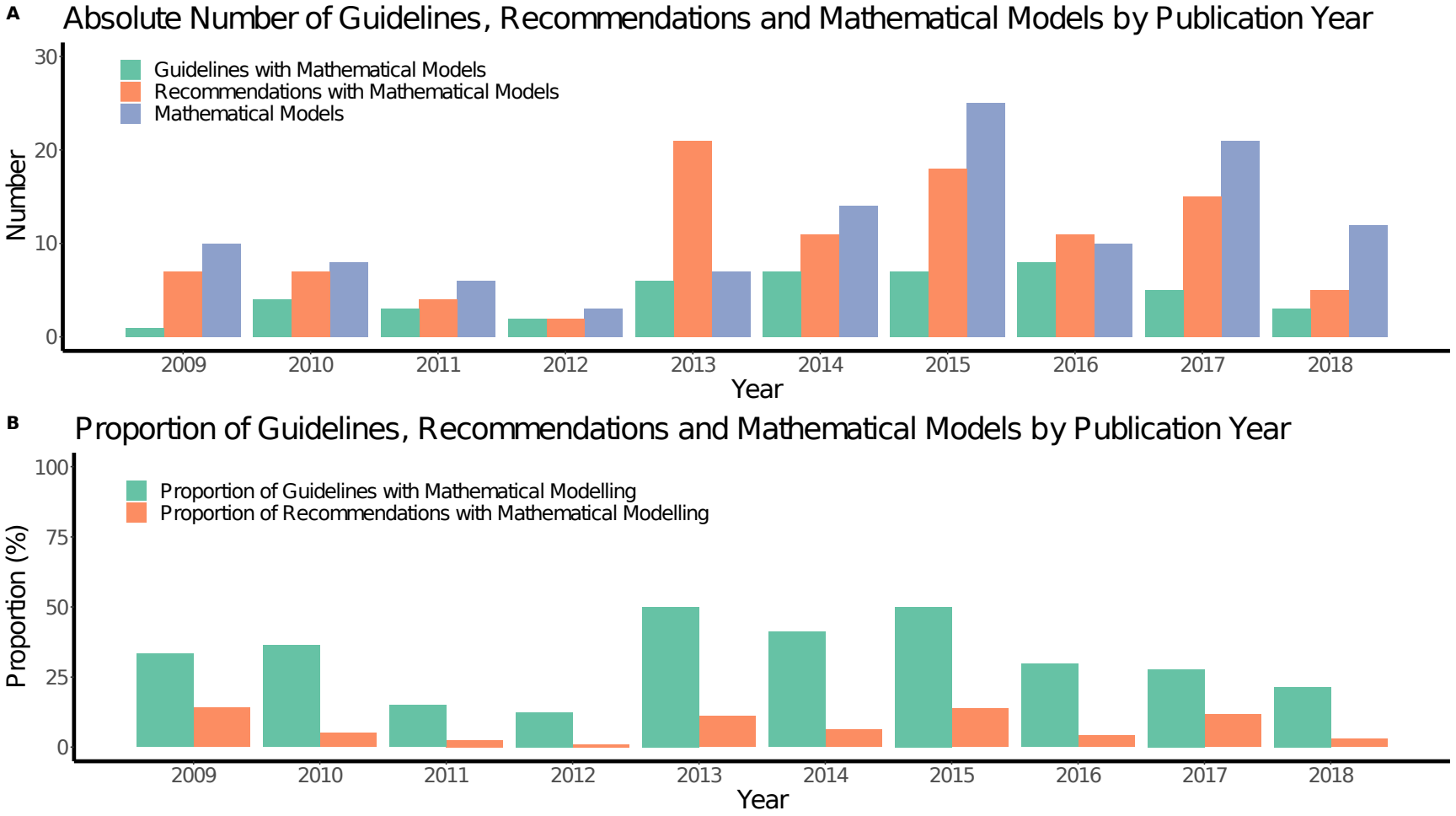

Figure 2: Temporal trends in the use of mathematical models in WHO guidelines and recommendations. We plotted WHO guidelines and guideline recommendations that use evidence from mathematical modeling studies for one or more recommendation by year of publication. (A) Absolute number of guidelines (green), guideline recommendations (orange), and distinct mathematical models cited (blue) by year; (B) Proportion of guidelines (green) and guideline recommendations (orange) informed by modeling, by year. We present the total number of guidelines and recommendations, by year, in Figure S1.

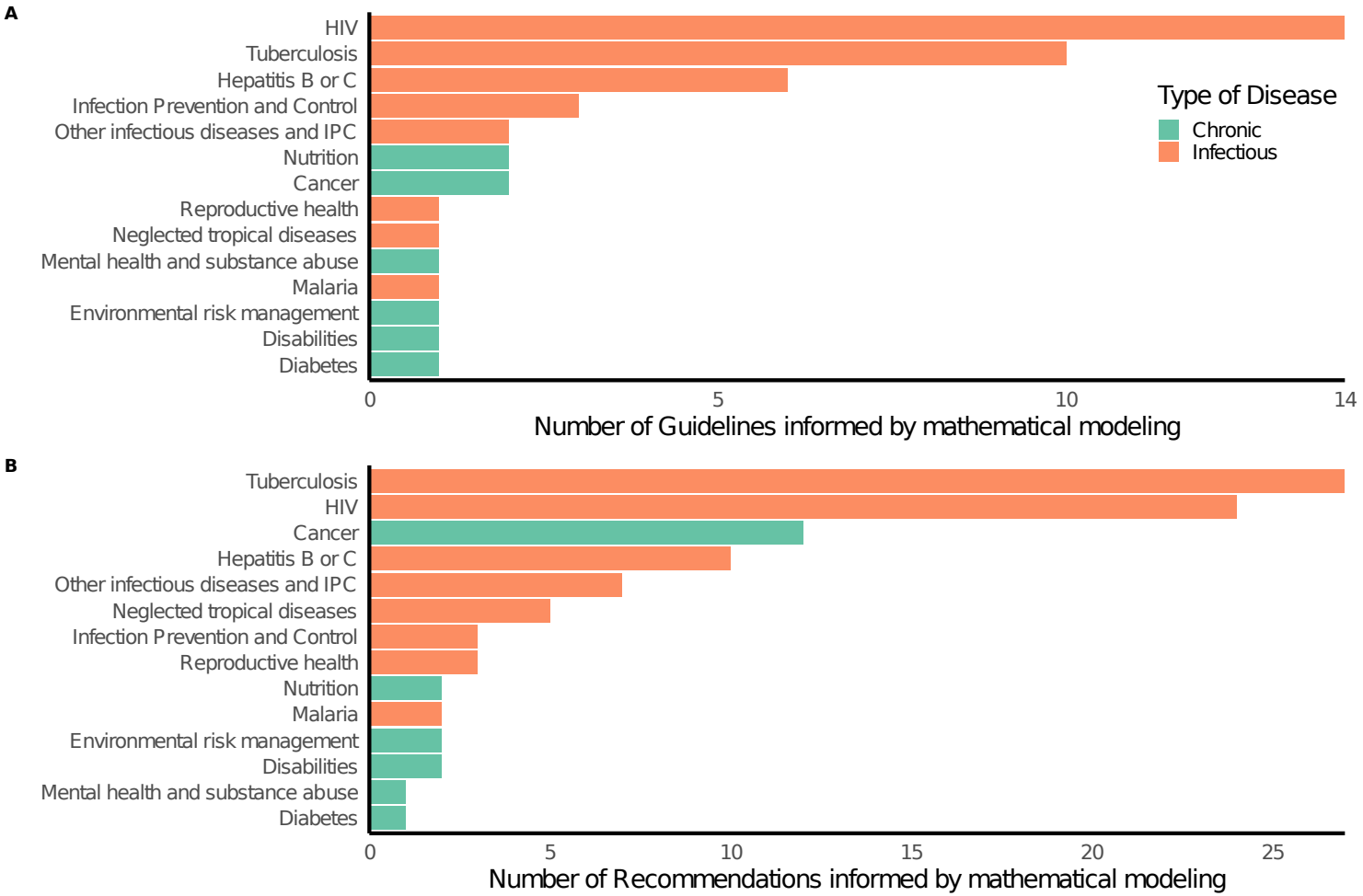

Figure 3: Summary of health topics for WHO guidelines and recommendations incorporating evidence from mathematical models. We plotted WHO guidelines (A) and guideline recommendations (B) that relied upon mathematical modeling evidence for one or more recommendation by health topic. Abbreviations: HIV- Human Immunodeficiency Virus, IPC- Infection Prevention and Control, TB- Tuberculosis, NCDs: Non-communicable diseases, Cancer: includes one guideline on cervical cancer and another guideline on mammography screening.

Acknowledgements

The authors appreciate research support from Dr Nicola Lo and Dr Anna Schoeni. Ms Marion Blacker provided administrative support.

Authorship contribution:

All authors had full access to all of the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis.

Study design- NCL, VIS, SLN, JL

Literature review and data collection- NCL, KA, PS, VIS, TB, SLN, JL

Data interpretation- NCL, KA, PS, VIS, SLN, JL

Wring of first draft- NCL

Contributed intellectual material and approved final draft- NCL, KA, PS, VIS, TB, SLN, JL

Declaration of interests:

SLN is an employee of WHO and oversees the quality of its guidelines; she is a member of GRADE Working Group. NCL and JAL report funding from the World Health Organization to support the current study. All other authors declare no conflicts of interest.

Financial disclosures:

JAL has received research grant funding from Merck Sharp & Dohme. NCL has received personal fees from the World Health Organization unrelated to the current study.

Funding/Support:

World Health Organization; Special Program for Research and Training in Tropical Diseases (TDR); Global Affairs Canada - Weapons Threat Reduction Program (Non-Proliferation and Security Threat Reduction Bureau)

Role of the Funding Organization or Sponsor:

The funding organizations had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; and preparation, review, or approval of the manuscript; or the decision to submit the manuscript for publication.

Disclaimer:

The authors alone are responsible for the views expressed in this article and they do not necessarily represent the views, decisions or policies of the institutions with which they are affiliated.

Previous presentations:

None.