A Framework for Synthesizing Intervention Evidence from Multiple Sources Into a Single Certainty of Evidence Rating: Methodological Developments from a US National Academies of Sciences, Engineering, and Medicine committee

Calonge, B.; Shekelle, G.P. ; Owens, D.K.; Teutsch, M.S.; Downey, A.; Brown, L.; Noyes, Jane

Research synthesis methods

DOI:
https://doi.org/10.1002/jrsm.1582

E-pub ahead of print: 20/06/2022

Received: 30 December 2021 | Revised: 27 April 2022 | Accepted: 11 May 2022

DOI: 10.1002/jrsm.1582

SPECIAL ISSUE PAPER

Research Synthesis Methods WILEY

A framework for synthesizing intervention evidence from multiple sources into a single certainty of evidence rating: Methodological developments from a US National Academies of Sciences, Engineering, and Medicine Committee

Ned Calonge[1,2] | Paul G. Shekelle[3,4] | Douglas K. Owens[5] | Steven Teutsch[6] | Autumn Downey[7] | Lisa Brown[7] | Jane Noyes[8]

[1]Department of Family Medicine, University of Colorado School of Medicine, Colorado School of Public Health, Aurora, Colorado, USA

[2]Department of Epidemiology, Colorado School of Public Health, Aurora, Colorado, USA

[3]General Internal Medicine Division, Greater Los Angeles Veterans Affairs Medical Center, Los Angeles, California, USA

[4]Department of Medicine, University of California, Los Angeles, California, USA

[5]Department of Health Policy, School of Medicine, Freeman-Spogli Institute for International Studies, Stanford University, Palo Alto, California, USA

[6]Department of Public Health, University of California, Los Angeles; and Senior Fellow, Leonard D. Schaeffer Center for Health, Policy and Economics, University of Southern California, California, Los Angeles, USA

[7]National Academies for Science, Engineering and Medicine, Washington, Washington, District of Columbia, USA

[8]Department of Health and Social Care Services Research and Child Health, School of Medical and Health Sciences, Bangor University, Bangor, Wales, UK

Correspondence
Ned Calonge, Department of Family Medicine, University of Colorado School of Medicine, 6506 E Progress Circle, Greenwood Village, CO 80111, USA.
Email: ned.calonge@gmail.com

Funding information
US Centers for Disease Control and Prevention

Abstract

Despite research investment and a growing body of diverse evidence there has been no comprehensive review and grading of evidence for public health emergency preparedness and response practices comparable to those in medicine and other public health fields. The National Academies of Sciences, Engineering, and Medicine convened an ad hoc committee to develop and use methods for grading and synthesizing diverse types of evidence to create a single certainty of intervention-related evidence to support recommendations for Public Health Emergency Preparedness and Response Research. A 13-step consensus building method was used. Experts were first canvassed in public meetings, and a comprehensive review of existing methods was undertaken. Although aspects of existing review methodologies and evidence grading systems were relevant, none adequately covered all requirements for this specific context. Starting with a desire to synthesize diverse sources of evidence not usually included in systematic reviews and using GRADE for assessing certainty and confidence in quantitative and qualitative evidence as the foundation, we developed a mixed-methods synthesis review and grading methodology that drew on (and in some cases adapted) those elements of existing frameworks and methods that were most applicable. Four topics were selected as test cases. The process was operationalized with a suite of method-specific reviews of diverse evidence types for each topic. Further consensus building was undertaken through stakeholder engagement and feedback The NASEM committee's GRADE adaption for mixed-methods reviews will further evolve over time and has yet to be endorsed by the GRADE working group.

KEYWORDS

evidence synthesis, GRADE, GRADE CERQual, mixed-methods, systematic review

Highlights

What Is Already Known
There is no method for reviewing evidence and assessing the certainty of mixed-methods systematic review findings in a single rating.

What Is New
Adapted from GRADE, a new method for synthesizing and assessing the certainty of mixed-methods systematic review findings in a single rating was developed using evidence-based principles.

Potential Impact for Research Synthesis Methods Readers outside the author's Field
The NASEM committee's GRADE adaption for mixed-methods reviews has potential for use in any other field where diverse sources of evidence contribute to the evidence base. For the first time Guideline panels and decision-makers can be presented with a single confidence of evidence rating for findings from mixed-methods syntheses of diverse sources of evidence to inform their decision-making.

1 | INTRODUCTION

As policy makers and practitioners have increasingly recognized the importance of having an evidence base to tackle complex challenges, there has been a growing movement among those who conduct systematic reviews and develop guidelines to embrace methods that take a complexity perspective and use multiple sources and types of evidence, requiring a shift away from a focus on simple, linear cause-and-effect models and related quantitative evidence, to "explore the ways in which interactions among components of an intervention or system give rise to dynamic and emergent behaviors".[1] p1 In the context of public health emergency preparedness and response, and in this paper, "intervention" refers to an action or actions taken in either preparation or response to a potential (future) or actual public health emergency. Multiple dimensions of intervention complexity may be considered in the evaluation of evidence, including complexity of intervention, pathway, population, context, feasibility, acceptability, cost, value, and implementation.[2] Reviewers and guideline developers have been developing and testing novel quantitative, qualitative, and mixed-methods for systematic reviews and evidence synthesis and grading to better capture complexity.[2–6] Mixed-methods research is defined by Pluye and Hong[7] as "a research approach in which a researcher integrates (a) qualitative and quantitative research questions, (b) qualitative research methods and quantitative research designs, (c) techniques for collecting and analyzing qualitative and quantitative evidence, and (d) qualitative findings and quantitative results." A mixed-methods synthesis can integrate quantitative, qualitative and mixed-methods evidence from primary studies. Evidence from mixed-methods primary studies is usually disaggregated into quantitative and qualitative evidence to synthesize in method-specific reviews and appraised using method-specific quality appraisal tools. Appraisal tools are available if wanting to appraise a mixed-methods primary study.

In addition to research-based evidence, both quantitative and qualitative, it is important in the context of public health emergency preparedness and response practice for the approach to evidence synthesis to make use of experiential evidence, whether quantitative or (most often) qualitative, such as that provided in case reports and after action reports from past response scenarios. After action reports are documents created by public health authorities and other response organizations following an emergency or exercise, primarily for the purposes of quality improvement.[8] They contain narrative descriptions of what was done but may also contain "lessons learned" (i.e., what was perceived to work well and not well) and recommendations for future responses. These evidence sources offer the potential for validation of research findings in practice settings, as well as improved understanding of context effects, trade-offs, and the range of implementation approaches or components for a given practice.

There are other categories of evidence that may also play a role in supporting guideline development such as modeling, mechanistic and parallel evidence. Modeling (specifically decision modeling) is a quantitative approach for simulating the benefits, harms, and costs of interventions when applied to a theoretic population/group of individuals. Modeling provides a formal methodological approach for extrapolating from available evidence to estimate what might happen in scenarios that could occur, such as a pandemic. "Mechanistic evidence", which can be informed by both quantitative and qualitative data, denotes relationships for which causality has been established—generally within other scientific fields, such as chemistry, biology, engineering, economics, and physics—and that can reasonably be applied to other contexts through mechanistic reasoning, defined in turn as "the inference from mechanisms to claims that an intervention produced" an outcome.[9 p434] "Parallel evidence" can be used to describe evidence on the effectiveness of similar practices from outside the context being examined. The consideration of supporting evidence from analogy (e.g., similar interventions or analogous contexts) was proposed by Sir Austin Bradford Hill[10] and has been resurrected in more recent discussions on evidence grading.[11] We relied on existing systematic reviews to contribute parallel evidence to the process.

Review methods for complex interventions and systems have thus focused on the integration of diverse and heterogeneous types of evidence from different types of studies as well as from different categories of evidence or "evidence streams". Qualitative and quantitative evidence may both contribute to understanding an intervention or practice and ultimately what works, necessitating synthesis approaches that combine these different types of evidence.[4,12] In some instances, guideline groups have synthesized across diverse evidence streams by mapping qualitative to quantitative findings or vice versa, so as to better understand the phenomenon of interest.[13–15] For example, the World Health Organization in their 2018 guideline on *Communicating risk in public health emergencies*, commissioned quantitative intervention effect reviews, qualitative evidence syntheses and gray literature reviews. Findings and results from individual reviews were integrated in an evidence to decision framework and presented as a final mixed-methods synthesis.[15]

While there have been several frameworks developed to assess the certainty and confidence in method-specific types of evidence, to date no one has operationally synthesized both quantitative and qualitative evidence inclusive of that from direct hypothesis-testing quantitative studies, qualitative studies, models, and mechanistic evidence and parallel evidence streams to arrive a single certainty of evidence rating for a finding. The evidence streams used by the committee in determining an overall certainty of evidence are summarized in Table 1.

Knowledge regarding evidence-based practice is critically needed in Public Health Emergency Preparedness and Response (PHEPR) given the mandate of the PHEPR system to mitigate the health, financial, and other impacts of public health emergencies. The PHEPR system, with its multifaceted mission to prevent, protect against, quickly respond to, and recover from public health emergencies,[16] is inherently complex and encompasses policies, organizations, and programs. This complexity also stems in part from the nature of public health emergencies, which are often unpredictable, may evolve rapidly, and are highly heterogeneous with respect to setting and type.[17] Setting is not limited to geographic location, but also encompasses the sociocultural and demographic environment, as well as the characteristics of the communities and the responding entities (e.g., organizational structure, managerial experience, staff capabilities, social trust, and other resources). PHEPR practices themselves may also be complex, featuring multiple interacting components that target multiple levels (e.g., individual, population, system), and with implementation that is often tailored to local conditions.[18]

The PHEPR system draws on different evidence types from a wide range of study designs and reports, from randomized controlled trials to after action reports, and the approach to evaluating the evidence needs to reflect that diversity. For all of these reasons, assessing the evidence base of public health emergency practices is an ideal opportunity to develop and operationalize methods for

TABLE 1 Summary of evidence streams, study designs and type of data

Evidence stream	Examples of study designs	Type(s) of data
Intervention studies		
Quantitative Comparative	Controlled experiment; observational study	Quantitative
Quantitative non-comparative	Single arm experiment; post event or intervention survey	Quantitative (note: surveys can also provide qualitative)
Qualitative	Phenomenological, generic qualitative research, grounded theory,	Qualitative
Mixed-methods	–	Quantitative and qualitative
Case report	–	Quantitative and qualitative
After action report	–	Quantitative and qualitative
Modeling studies	Decision models, simulations	Quantitative
Mechanistic	Post-event investigations	Qualitative
Parallel	Systematic evidence reviews	Quantitative

synthesizing evidence across diverse streams and to develop a single certainty of evidence rating for a finding. The National Academies of Science, Engineering and Medicine (NASEM) committee on Evidence-Based Practices for Public Health Emergency and Response set about developing and using methods for grading and synthesizing diverse types of evidence to create a single certainty of intervention-related evidence to support evidence-based recommendations for PHEPR practice.[19]

2 | METHODS

The committee included 20 methodological and subject specific experts, and administrative oversight. Nine additional subject matter experts and one methodological consultant advised the committee. Standard methods for tool development were used including evidence and expert review, methodological development by consensus and feedback from the wider community of stakeholders including the funder (Centers for Disease Control and Prevention), evidence users, practitioners and methodologists, including representation from the GRADE working group.[20,21] The method for grading mixed-methods evidence developed by the NASEM committee on Evidence-based Practices for Public Health Emergency and Response ("the committee") involved adopting and/or adapting existing established method-specific methods for searching and assessing qualitative and quantitative evidence from different types of studies and reports (quantitative comparative studies, quantitative non-comparative studies, qualitative studies, mixed-methods studies, descriptive surveys, case reports and after action reports) and then a novel method for synthesizing findings across different evidence streams (intervention study evidence, modeling evidence, mechanistic evidence and parallel evidence). The primary literature search strategy was restricted to emergency preparedness and response interventions in public health settings only within a specified publication timeframe and inclusion was systematic based on title and abstract review by two independent reviewers. The steps followed are outlined in Table 2.

The committee heard from experts in the Grading of Recommendations Assessment, Development, and Evaluation (GRADE), which is used in WHO guidelines,[14,21] the Community Preventive Services Task Force,[22] the US Preventive Services Task Force (USPSTF),[23] the National Aeronautics and Space Administration Integrated Medical model,[24] the Clearinghouse for Labor Evaluation and Research (CLEAR),[25] the What Works Clearinghouse,[26] the Evaluation of Genomic Applications in Practice and Prevention (EGAPP),[27] the National Highway Traffic Safety Administration Countermeasures that Work,[28] and considered other methods or frameworks for establishing causality (such as the criteria developed by Bradford Hill). Having concluded that, while each evidence grading system had its strengths, no one grading system captured all of what the committee felt was necessary for the requirements of this specific mixed-methods context. The committee set out to develop its own method, an adaption of GRADE* for mixed-methods evidence. *The adaption has not yet been endorsed by the GRADE working group.

The two overarching principles were:

1. To be grounded to the greatest extent possible in existing frameworks and evidence grading systems.
2. To be sufficiently flexible to accommodate evidence from all the different streams considered important by PHEPR stakeholders.

TABLE 2 Steps and processes in the consensus development method

Step	Process
1	Evidence review of current evidence grading methods
2	Public meetings to hear from experts on existing evidence evaluation and grading methods
3	Consensus development of an initial framework to grade evidence of different types in a single final rating
4	Select the review topic, considering published literature on gaps/priorities and stakeholder input.
5	Develop the analytic framework and key review questions.
6	Conduct a search of the peer-reviewed and gray literature and solicit papers from stakeholders.
7	Apply inclusion and exclusion criteria.
8	Separate evidence into methodological streams (quantitative studies, including comparative, noncomparative, and modeling studies and descriptive surveys; qualitative studies; after action reports and case reports) and extract data.
9	Apply/adapt existing tools for quality assessment of individual studies based on study design.
10	Synthesize the body of evidence within methodological streams and apply an appropriate grading framework (Grading of Recommendations Assessment, Development and Evaluation [GRADE] for the body of quantitative research studies and GRADE-Confidence in the Evidence from Reviews of Qualitative Research [GRADE-CERQual] for the body of qualitative studies to assess the certainty of the evidence [certainty of evidence]/confidence in the findings, respectively).
11	Consider evidence of effect from other streams (e.g., modeling, mechanistic, qualitative evidence) and support for or discordance with findings from quantitative research studies to determine the final certainty of evidence.
12	Draft Recommendations taking account of the overall certainty of evidence.
13	Disseminate methods and findings for feedback to consolidate consensus.

In keeping with the first principle, the committee adopted the approach of the USPSTF and the Community Preventive Services Task Force by a priori specifying the analytic framework for each key question—a visual depiction of the pathway between the interventions of interest and the outcomes of interest. The analytic frameworks then were used to identify the questions for the literature searches. These in turn were then performed using methods standard to each evidence stream. Thus, quantitative comparative studies were assessed using standard risk of bias tools[29] and then had the certainty of evidence determined using GRADE,[21] qualitative studies had their methodological limitations assessed using the Critical Appraisal Skills Programme (CASP) tool[30] and then had confidence in the evidence assessed using GRADE-CERQual,[31] modeling studies were assessed by an expert in modeling methodology. Modeling studies were not included in the bodies of evidence assessed with the GRADE domains but were considered in the overall certainty of evidence determination. We reviewed modeling studies to assess their methodological approach, data sources, relevance to key questions, and implications for public health practice. Based on this review, we selected the most relevant studies for intensive review by a highly experienced modeling expert who conducted a detailed assessment of the strengths and limitations of the analyses. Descriptive case reports do not fit any specific analytic study design and generally report few details concerning methods, and thus are not amenable to quality assessment using tools designed for research studies. Case reports and after action reports were categorized as "high" or "low" priority using the significance criterion of the AACODS (authority, accuracy, coverage, objectivity, data, significance) checklist.[32] An appraisal tool for evaluating the methodological rigor of after action reports published in 2019 (ECDC, 2018) was applied.[33] Figure 1 depicts how the literature review and assessment of studies in different methodologic streams was performed.

In keeping with the second principle, the method had to consider more than just published evidence specifically about PHEPR; public health stakeholders also considered that mechanistic evidence and parallel evidence were important in their decision-making. Mechanistic evidence does not have a universally agreed-upon definition but can be considered to be relationships for which causality has been established—often but not exclusively from other scientific fields, such as chemistry, biology, economics, and physics—which can then reasonably be applied to the specific context. Mechanistic evidence is often used, for example, in National Transportation Safety Board determinations of what caused air traffic accidents, such as the mid-flight explosion of TWA Flight 800.[34] An example in the field of Public Health Emergency Preparedness is the recommendation to site hospital auxiliary generators above the highest anticipated water line in flood prone areas, as opposed to siting such generators in the hospital basement. The

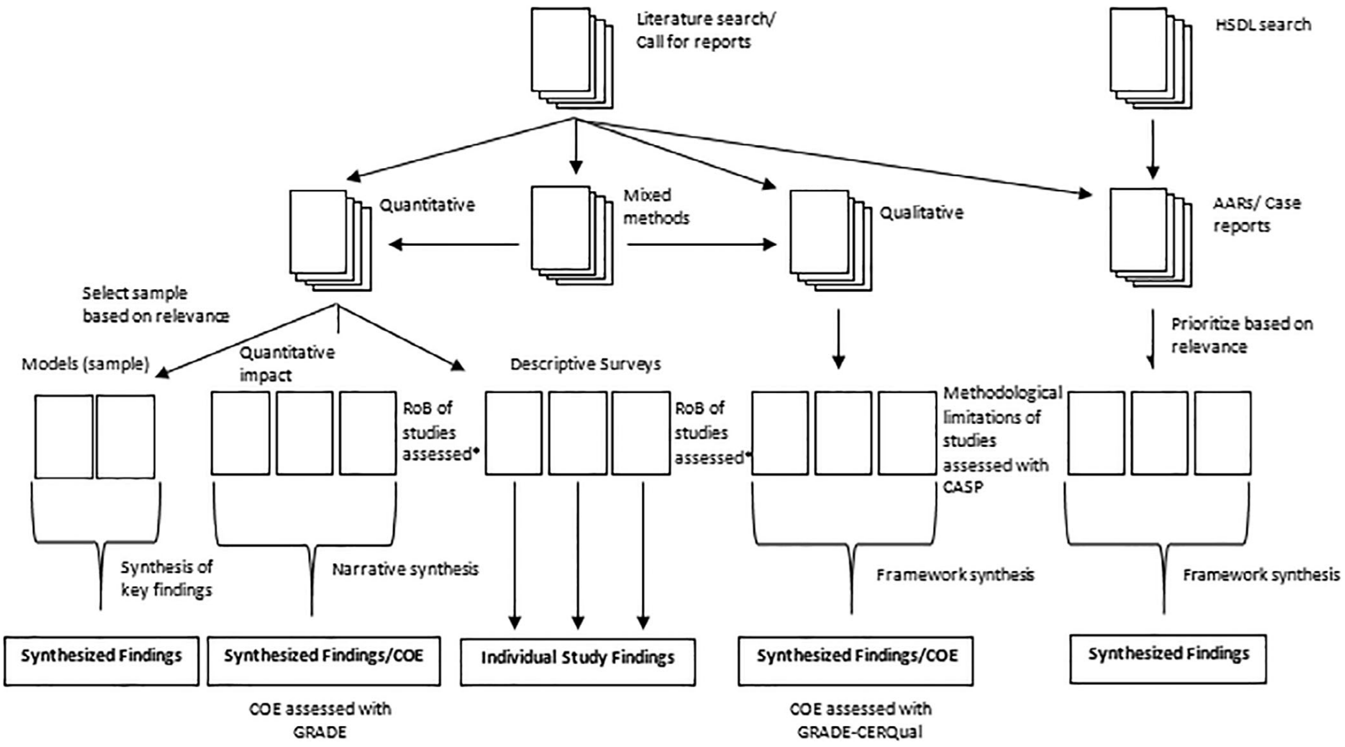

FIGURE 1 Classification and consolidation of studies into methodological streams. Abbreviations: AAR, after action report; CASP, Critical Appraisal Skills Programme; certainty of evidence, certainty of the evidence; GRADE, Grading of Recommendations Assessment, Development and Evaluation; HSDL, Homeland Security Digital Library; RoB, risk of bias.*Risk of bias assessment tools were developed by adapting existing tools and/or published methods

mechanistic reasoning in this example is that water seeks its own level, and that flooded generators can short circuit and not work. Therefore, siting the generator above the high water line should reduce the possibility a hospital will go powerless during a flood.

Parallel evidence refers to evidence about the effectiveness of similar practices from outside the context of interest, in this case PHEPR. It is related to the Bradford Hill[9] criteria of analogy, and somewhat related to the GRADE domain of indirectness. But whereas indirectness in GRADE normally refers to evidence about the same intervention but using a different outcome (such as an intermediate process outcome rather than a health outcome) or a different population (such as middle-aged adults when the population of interest is over age 65 adults) and is always used to downgrade the certainty of evidence, as used by the committee (and consistent with how Bradford Hill[9] proposed using evidence by analogy) parallel evidence can be about interventions that may not be identical to the intervention of interest in addition to being about other populations or outcomes, and may be used to upgrade the certainty of evidence. Parallel evidence in medicine is what allows us to conclude with high certainty that oral steroids relieve pain in acute gout, because we know that intra-articular steroids relieve pain in acute gout and oral steroids relieve symptoms in acute inflammatory exacerbations of almost any kind.

The committee then applied and further improved the NASEM committee's GRADE adaption for mixed-methods evidence by using it to evaluate four PHEPR practices, purposively chosen to represent the diversity of practices in PHEPR. The four PHEPR practices, chosen using a formal method explained in further detail in the report,[19] were (1) engaging with and training community-based partners to improve the outcomes of at-risk populations after public health emergencies, (2) activating a public health emergency operations center, (3) communicating public health alerts and guidance with technical audiences during a public health emergency, and (4) implementing quarantine to reduce or stop the spread of a contagious disease. Experienced review groups were commissioned to search and critically appraise the literature in the different data streams.[19]

At the end of this stage, for each key question the committee had before it the results of the identification and critical appraisal of evidence from quantitative comparative studies, qualitative studies, case reports, after action reports, modeling studies, parallel evidence, and mechanistic evidence. Figure 2 depicts how these were synthesized into a single certainty of evidence rating.

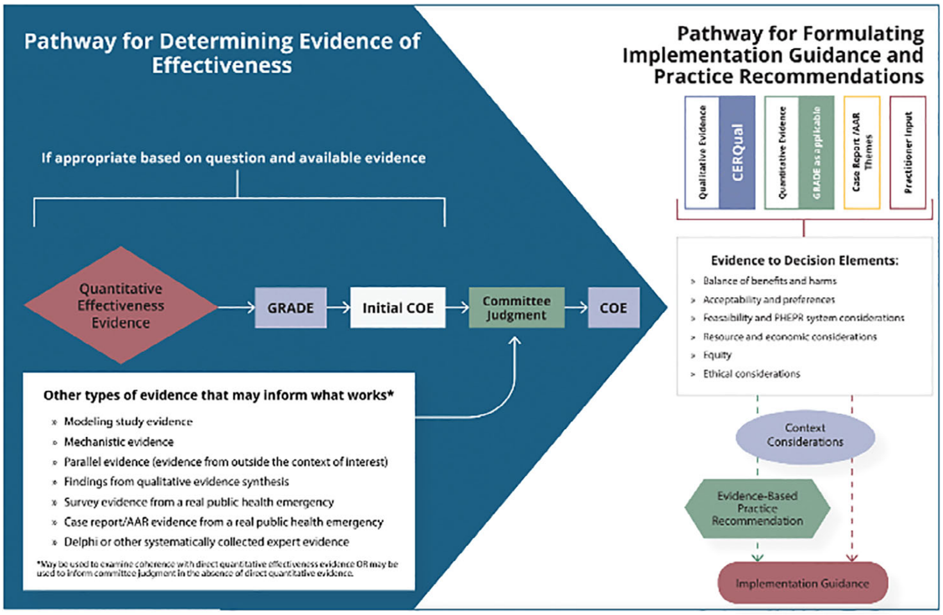
Pathway for Determining Evidence of Effectiveness
If appropriate based on question and available evidence
Quantitative Effectiveness Evidence
GRADE
Initial COE
Committee Judgment
COE
Other types of evidence that may inform what works*
Modeling study evidence
Mechanistic evidence
Parallel evidence (evidence from outside the context of interest)
Findings from qualitative evidence synthesis
Survey evidence from a real public health emergency
Case report/AAR evidence from a real public health emergency
Delphi or other systematically collected expert evidence
*May be used to examine coherence with direct quantitative effectiveness evidence OR may be used to inform committee judgment in the absence of direct quantitative evidence.
Pathway for Formulating Implementation Guidance and Practice Recommendations
Qualitative Evidence
CERQual
Quantitative Evidence
GRADE as applicable
Case Report /AAR Themes
Practitioner Input
Evidence to Decision Elements:
Balance of benefits and harms
Acceptability and preferences
Feasibility and PHEPR system considerations
Resource and economic considerations
Equity
Ethical considerations
Context Considerations
Evidence-Based Practice Recommendation
Implementation Guidance

FIGURE 2 Framework for integrating evidence to inform recommendation and guidance development for PHEPR practices. The framework depicts two interconnected pathways for evaluating evidence for PHEPR practices. The lefthand panel (blue) shows our process for integrating evidence from quantitative impact studies with other evidence that may inform what works to determine the certainty of the evidence (certainty of evidence) of effectiveness for a given outcome. The certainty of evidence (for all relevant outcomes) feeds into the righthand panel (white), which shows the pathways for integrating diverse evidence for various elements (evidence to decision elements) that, along with context considerations, may inform the formulation of evidence-based practice recommendations and implementation guidance. In cases in which the review is focused on implementation and not on determining the effectiveness of a practice, it is possible to follow the pathway depicted in the righthand panel without assessing the certainty of evidence as shown in the left hand panel [Colour figure can be viewed at wileyonlinelibrary.com]

This started with an assessment of the quantitative comparative studies, using GRADE to determine an initial certainty of evidence. GRADE certainty of evidence ratings are based on assessments in five domains (risk of bias, indirectness, imprecision, inconsistency, and publication bias) and have additional instructions for upgrading and downgrading (see Table 3). GRADE defines certainty of evidence as follows:

- High certainty—We are very confident that, in some contexts, there are important effects (benefits or harms). Further research is very unlikely to change our conclusion.
- Moderate certainty—We are moderately confident that, in some contexts, there are important effects, but there is a possibility that there is no effect. Further research is likely to have an important impact on our confidence and could alter the conclusion.
- Low certainty—Our confidence that there are important effects is limited. Further research is very likely to have an important impact on our confidence and is likely to change the conclusion.
- Very low certainty—We do not know whether the intervention has an important effect.

For qualitative studies, the committee used the standard framework synthesis method,[35] which employs an iterative deductive and inductive process, to analyze and synthesize the findings.[19] GRADE-CERQual[31] was used to assess the confidence in synthesized qualitative findings. CERQual provides a systematic and transparent framework for assessing confidence in individual review findings, based on consideration of four components:

- methodological limitations—the extent to which there are concerns about the design or conduct of the primary studies that contributed evidence to an individual review finding.
- coherence—an assessment of how clear and compelling the fit is between the data from the primary studies and a review finding that synthesizes those data.
- adequacy of data—an overall determination of the degree of richness and quantity of data supporting a review finding, and
- relevance—the extent to which the body of evidence from the primary studies supporting a review finding is applicable to the context (perspective or population, phenomenon of interest, setting) specified in the review question.

TABLE 3 Grading of recommendations assessment, development and evaluation (GRADE) evidence evaluation domains

GRADE domains for assessing certainty of the evidence
Downgrading domains
• *Risk of bias*—the potential for limitations in the study design and execution to influence estimates of the intervention effect. Risk-of-bias assessments for all individual studies are included in the body of evidence.
• *Indirectness*—considers whether the available evidence differs from the target of interest, including differences in population, interventions, outcome measures (e.g., use of surrogate outcomes removed in the putative causal pathway from important endpoints), and comparison groups.
• *Imprecision*—when study results include relatively few participants/events and thus have a wide confidence interval (CI) around the estimate of effect.
• *Inconsistency*—unexplained heterogeneity of results across studies.
• *Publication bias*—systematic underestimation or overestimation of the underlying beneficial or harmful effect due to the selective publication of studies.
Upgrading domains
• *Large effect*—considers whether an effect is large enough that it cannot have occurred solely as a result of bias from potential confounding factors.
• *Dose–response gradient*—refers to an observation of progressively larger effect with greater exposure to the intervention.
• *Plausible residual confounding*—if confounding is likely to work counter to what the evidence demonstrates (would decrease an apparent intervention effect or would create a spurious effect when results suggest no effect), it may confer greater confidence in the evidence.

TABLE 4 Matrix with the generalized approach to determine the certainty of the Evidence (certainty of evidence)

Certainty of evidence Decision	Criteria
No change in certainty of evidence	Did not upgrade based solely on evidence from case reports, surveys, supportive evidence from modeling evidence, or low-confidence findings from qualitative evidence synthesis. Did not upgrade for supportive parallel evidence when direct evidence (from the PHEPR context) was available that resulted in low or moderate initial certainty of evidence. Did not upgrade if evidence raised concerns about potential harmful/undesirable effects.
Upgraded certainty of evidence one level	Required very supportive mechanistic or modeling evidence or high-confidence findings from qualitative evidence synthesis.
Upgraded certainty of evidence two levels	Required a combination of supportive (or very supportive) findings from mechanistic, modeling, or qualitative evidence.
Downgraded certainty of evidence	Although we did not encounter this scenario, evidence of harmful/undesirable effects could warrant downgrading the initial certainty of evidence.

Based on these ratings, each synthesized finding was then assigned an overall assessment as follows:

- High confidence—It is highly likely that the finding is a representation of the phenomenon.
- Moderate confidence—It is likely that the finding is a representation of the phenomenon.
- Low confidence—It is possible that the finding is a representation of the phenomenon.
- Very low confidence—It was not clear whether the finding is a representation of the phenomenon.

The committee then applied the GRADE adaption for mixed-methods evidence by using the criteria in Table 4 to upgrade or downgrading certainty, based on the results of the assessments from the other evidence streams. As with GRADE and other systematic review and guideline development processes, there is no algorithm or mathematical model for determining the certainty of evidence (such as based on the number and quality of included studies). The assessment of the certainty of evidence is based on the judgment of the evaluators. In some cases, a single high-quality study may provide a high certainty of evidence, while in others, having multiple RCTs with consistent effects could yield a lower certainty of evidence (e.g., due to indirectness). For transparency, the committee worked to be clear about the rationale for up- and/or downgrading decisions and the ultimate certainty of evidence rating. Each additional source of evidence was judged to be supportive, very supportive, inconclusive (no conclusion can be drawn regarding coherence because either results are mixed or the data are insufficient), or unsupportive (discordant with the findings from quantitative impact research studies). The distinction between supportive and very supportive evidence was based on the magnitude of the reported effect and the directness of its application to the question and outcome of interest. Mechanistic evidence, which does not lend itself to an assessment of magnitude of effect, was determined to be supportive or very supportive based on the counterfactual (i.e., how likely it is that an alternative explanation accounts for the observed effect that has been attributed to a specified mechanism of action). While an observed reduction in disease transmission may reasonably be attributed to quarantine based on its mechanism (i.e., separating individuals at risk of becoming infectious from susceptible populations), other factors (e.g., seasonal effects related to temperature and humidity) may actually be responsible for the reduced spread. In contrast, mechanistic evidence regarding the impact of congregate quarantine was deemed very supportive as there is no good alternative explanation for why infections would increase among those quarantined in the congregate setting. A global judgment was made as to whether there was sufficient supportive or unsupportive evidence to warrant up- or downgrading the initial certainty of evidence. These initial ratings were performed by the committee chair, in consultation with NASEM staff members. These ratings were presented and discussed, one-at-a-time, to the full committee during a face-to-face meeting, at which time revisions were made with full committee input.

2.1 | Further consensus building activities

Supporting Information, Data S1, provides details of additional dissemination and consensus building activities through which the methods and recommendations were presented and feedback invited. Supporting Information, Data S2, outlines the impact and global reach of the consensus report of the methodological development and practice recommendations as of 21.12.21.

3 | RESULTS

Table 5 shows the yield of the literature searches for each of the four PHEPR practices assessed, by the type of evidence found. For one topic (activating a public health emergency operations center) no quantitative evidence was identified, and the committee did not believe it could use the NASEM committee's GRADE adaption for assessing certainty of evidence without at least some quantitative evidence. Because of this, that topic was dropped from the remainder of the analysis. As seen in Table 5, there were substantial amounts of additional evidence in data streams other than quantitative comparative studies, but the types and amount differed across topics. So, for example, there were many modeling studies about the effects of quarantine, and none about any of the other topics. Communicating alerts and guidance with technical audiences had relatively few qualitative studies compared to the other topics. Parallel evidence was common for the topic about engaging and training community-based partners so we used this as a test case. We did not attempt to identify parallel evidence for other topics as the committee concluded it would not be applicable or useful in the other test cases.

Table 6 shows the results of the synthesis of evidence across the evidence streams into a single certainty of evidence rating using the NASEM committee's GRADE adaption for mixed-methods evidence. As noted above,

TABLE 5 Yield of literature searches for studies for four PHEPR topics

	Number of Studies			
Evidence type	**Engaging with and training community-based partners**	**Activating public health emergency operations**	**Communicating public health alerts and guidance with technical audiences**	**Quarantine to reduce or stop the spread of a contagious disease**
Quantitative Comparative	7	0	2	9
Quantitative Noncomparative (postintervention measure only)	4	0	0	4
Qualitative	23	21	8	16
Modeling	0	0	0	12
Descriptive Surveys	7	1	8	13
Case reports	15	29	12	28
After Action Reports	N/A	35	29	N/A
Mechanistic	N/A	N/A	N/A	Yes
Parallel (systematic reviews)	13	N/A	N/A	N/A

TABLE 6 Initial and final certainty of evidence for conclusions about evidence for three topics

Finding	# of Quantitative comparative studies	Original certainty of evidence using GRADE	Other sources of evidence					Final certainty of evidence	Change in certainty of evidence
			Qualitative evidence	Modeling evidence	Parallel evidence	Mechanistic evidence	Case report/after action report evidence		
Effect of culturally tailoring programs on knowledge of partners	3	Low	–	–	Supportive	–	–	Low	0
Effect of culturally tailoring programs on attitudes of partners	1	Very low	–	–	–	–	–	Very low	0
Effect of culturally tailoring programs on planning by partners	2	Very low	–	–	–	–	–	Very low	0
Effect of culturally tailoring programs on knowledge of at-risk populations	4	Moderate	–	–	Supportive	–	–	Moderate	0
Effect of culturally tailoring programs on attitudes of at-risk populations	3	Low	–	–	–	–	–	Low	0
Effect of culturally tailoring programs on behaviors of at-risk populations	4	Moderate	–	–	Supportive	–	–	Moderate	0
Effect of partner engagement on preparedness outreach activities	1	Very low	–	–	–	–	Supportive	Very low	0
Effect of partner engagement on coalitions addressing resilience	1	Very low	–	–	–	–	–	Very low	0

TABLE 6 (Continued)

Finding	# of Quantitative comparative studies	Original certainty of evidence using GRADE	Other sources of evidence					Final certainty of evidence	Change in certainty of evidence
			Qualitative evidence	Modeling evidence	Parallel evidence	Mechanistic evidence	Case report/after action report evidence		
Effect of electronic messaging on improved technical audience awareness	2	Moderate	–	–	–	–	>3	Moderate	0
Effect of electronic messaging on improved technical audience use of guidance	1	Very low	–	–	–	–	>3	Very low	0
Effect of quarantine on reduced transmission	3	Low	–	Supportive	–	Supportive	Supportive	**High**	**+2**
Effect of quarantine on reduced time to symptom onset	1	Very low	–	–	–	Supportive	Supportive	**Low**	**+1**
Effect of congregate quarantine on increased risk of infection	2	Moderate	–	–	–	Very supportive	1	**High**	**+1**
Effect of quarantine on risk of psychological harms	6	Low	Very supportive	–	–	–	Supportive	**Moderate**	**+1**
Quarantine and risk of financial hardship	2	Low	Very supportive	–	–	Very supportive	–	**High**	**+2**
Health-promoting leadership and reduced depression and anxiety in quarantined persons	1	Very low	–	–	–	–	–	Very low	0

(Continues)

TABLE 6 (Continued)

Finding	# of Quantitative comparative studies	Original certainty of evidence using GRADE	Other sources of evidence					Final certainty of evidence	Change in certainty of evidence
			Qualitative evidence	Modeling evidence	Parallel evidence	Mechanistic evidence	Case report/after action report evidence		
Risk communication and messaging about employee leave and improved adherence to quarantine	1	Low	Very supportive	–	–	–	–	**Moderate**	**+1**

The final certainty of evidence and change in certainty of evidence are bolded.

this was done only for the topics that had quantitative comparative evidence, since this is required as the initial step of the process. For the majority of the conclusions across all three topics, there was no difference between the initial certainty of evidence rating, based on the GRADE evaluation of the quantitative comparative evidence, and the final certainty of evidence rating using the NASEM committee's GRADE adaption for mixed-methods evidence, either because there was no other evidence considered or because the other evidence was insufficient to change the initial certainty of evidence, using the criteria in Table 4. For one topic—quarantine—most initial certainty of evidence ratings were changed by the inclusion of other evidence when the NASEM committee's GRADE adaption for mixed-methods evidence was applied. In most of these, the change in certainty of evidence was one level, for example from an initial certainty of evidence of Low to a final certainty of evidence of Moderate. In two conclusions, the change was two levels, and we discuss them in more detail here, to explain the reasoning for these changes.

Example #1: Effect of quarantine on overall disease transmission in the community: the role of mechanistic evidence and modeling evidence. Change in certainty of evidence from low to high.

The literature search yielded three quantitative comparative studies: a controlled before-and-after study of home quarantine versus standard operating procedures for control of influenza H1N1 at two Japanese auto factories; a retrospective analysis of the effect of quarantine on the 2003 SARS outbreak in Toronto; and a retrospective comparison of quarantine versus an undescribed sample of people meeting quarantine criteria but not quarantined during a measles outbreak in Switzerland. All studies had methodological limitations, and the body of evidence was judged to have serious limitations due to risk of bias and due to indirectness, yielding an initial certainty of evidence of low. However, there is also a mechanistic rationale to support that quarantine works—for example the germ theory of disease is predicated on the transmission of the germ from one host to another, and the 18th century practice of requiring arriving ships to quarantine at anchor for a period of time before allowing any disembarkation. Added to this are the results of numerous modeling studies that identify the circumstances of an infection that make quarantine more or less effective (such as the reproductive number [R_0], a short incubation period, and a relatively short asymptomatic period). These additional considerations resulted in a determination that the certainty of evidence is high, (applying the NASEM committee's GRADE adaption for mixed-methods evidence) that quarantine can reduce overall disease transmission in the community in certain circumstances (such as the ones identified by the modeling studies).

Example #2: Effect of quarantine on financial hardship for those quarantined: the role of mechanistic evidence and qualitative evidence. Change in certainty of evidence from low to high.

With regard to the potential harms of quarantine, the literature search identified two quantitative studies, both of them cross-sectional surveys of persons after epidemics, one of which was the 2003 SARS outbreak in Toronto and the other after the H1N1 influenza epidemic in Australia. Both surveys found that being placed in quarantine led to financial difficulties in quarantined individuals. This body of evidence was judged to be at serious risk of bias, yielding an initial certainty of evidence of low. However, there is also a strong mechanistic rationale to support that quarantine may cause financial hardship: some of the people being quarantined have jobs, and some of the people with jobs will not be able to work them while in quarantine, and some of these people not working their jobs will get reduced or zero pay, and thus face financial hardship. This finding is also strongly supported by high confidence in qualitative evidence from five studies, which found that people were often placed in quarantine with little advance notice, which affected their employment status and resulted in loss of income. These additional considerations resulted in a determination that the certainty of evidence is high (applying the NASEM committee's GRADE adaption for mixed-methods evidence) that a potential harm of quarantine is financial harm for quarantined individuals.

4 | DISCUSSION

We developed a method for rating the overall certainty of evidence that could accommodate diverse types of evidence including randomized trials, observational studies, qualitative evidence, mechanistic evidence, modeling studies, after action reports, and case studies. We are unaware of any method for doing such an assessment that existed prior to our work. We subsequently brought these streams of evidence together, along with parallel and mechanistic evidence where appropriate, in a single integrated mixed-methods synthesis using a logic model as the analytical framework for integration. We adopted as the foundation for our layered grading approach the widely used GRADE method for evaluating quantitative evidence of effectiveness and the GRADE-CERQual method for assessing synthesized qualitative findings. The GRADE approach is used in WHO guidelines and to date GRADE and GRADE CERQual assessments have not been integrated into an overall assessment. However, our NASEM committee's GRADE adaption for mixed-methods evidence with an integrated certainty of evidence assessment described here went beyond the GRADE approach. Consequently, the use of GRADE and GRADE-CERQual gives reviewers access to widely used evidence evaluation tools that are regularly updated.

A key feature of our methodology to develop a single certainty of evidence rating for a finding is that it had to accommodate the broad range of evidence relevant to PHEPR decision making and other similar fields. Incorporating evidence from evidence streams outside those normally considered by GRADE was necessary for our results to have face validity with the intended target audience and improved the assessment by being inclusive of other types of recognized evidence. Such evidence includes RCTs, nonrandomized experimental studies, case reports, modeling studies, and descriptive surveys, as well as mechanistic evidence and parallel evidence from other fields. In assessing certainty of evidence in the four PHEPR topics, we experienced challenges applying some of the GRADE domains. GRADE is most suitable for discrete interventions as is typical in clinical trials, but perhaps less so for more complex areas where context and the effect of multiple interventions are prominent study characteristics[36] We judged that it would not be conceptually appropriate to assume that an effect size existed independent of context and implementation fidelity. Further consideration of potential modifications to GRADE or of alternative rating schemes that provide more emphasis on non-RCT methods is warranted.

Although it is common for evidence review groups to exclude studies based on study design or methodological limitations in execution, we instead considered the appropriateness of the study design and the quality of execution as they related to the ability to address a specific review question. For example, qualitative research methods were considered superior to quantitative methods for certain tasks, such as describing the lived experiences of people placed under quarantine or exploring the ways in which multiple factors coalesce or conflict in the minds of decision makers choosing whether to implement an emergency operations center. Since much learning about what works and considerations for implementation accumulates through experience, it was important for the mixed-methods synthesis approach to accommodate experiential evidence, such as case reports and after action reports, so as to corroborate research findings in the NASEM committee's GRADE adaption for mixed-methods evidence certainty of evidence determination and help to explain differences in outcomes in practice settings (for example, by illustrating differences in feasibility or acceptability across settings). However, integrating evidence from after action reports and case reports presented its own challenges since these types of reports rarely include clear outcome measures or clearly elucidated cause-effect relationships. Moreover, such evidence, even when derived in accordance with high methodological standards, is subject to higher risk of bias compared with evidence from randomized controlled trials. We attempted to mitigate these risks by ensuring that the methods used to assess the quality of evidence were suited not just to the type of evidence being reviewed but also to the purpose to which that evidence was to be put, rather than holding every study to the same set of evaluative criteria. For example, the quality threshold for applying evidence to an assessment of *acceptability* differed from that for assessing *effectiveness*.

We also used a consensus-based judgment approach for the NASEM committee's GRADE adaption for mixed-methods evidence, which allowed the evaluators flexibility in the certainty of evidence determination process, a potential limitation is poor interrater reliability (i.e., others could arrive at different judgments given the same set of diverse evidence).

An added challenge for our work was the lack of existing quality assessment and grading methods for bodies of descriptive surveys and case reports/after action reports. For example, the appraisal tool for evaluating the methodological rigor of after action reports published in 2019[33] was not useful in selecting reports to include in the synthesis of after action reports and case reports because of the generally low scores for the majority of reports captured in the search. With improvements in the methodological rigor of after action reports, however, such tools could be helpful in selecting high-quality after action reports for inclusion in future evidence reviews. Consequently, some of the evidence streams used were synthesized and graded, while others were not. Other groups have adapted the GRADE and GRADE-CERQual methods for these evidence types,[13] but in the absence of methods for integrating the assessments to generate a composite rating, we chose not to grade bodies of descriptive surveys and case reports/after action reports. Given these gaps in evidence review methods, we took a pragmatic approach to integrating the diverse evidence types that were captured in its reviews, as described above. However, as the methodological science behind mixed-methods synthesis continues to evolve, it will be important to update the methods. Thus, the methods presented here should not be viewed as the final word in how PHEPR topics should be systematically assessed, but rather the starting point to be built on in future efforts.

Although models have been incorporated into past evidence reviews, such as the Community Guide review of school closure to reduce transmission of pandemic influenza,[37] this remains an active area of methodological development and is also an intensive process. Consequently, we undertook only a limited analysis. As methods for review and integration of modeling evidence are refined, the methodology applied will need to be updated.

The use of mechanistic evidence in evidence syntheses is uncommon, although evidence of biological mechanisms of action is increasingly being incorporated into reviews, for example, on pharmacological and toxicological topics. This is another area requiring further methodological development, one that would benefit from the efforts of a future guidelines development group to further develop and refine the definition and test the mechanistic upgrading assumptions.

For most of the interventions in our test cases, the inclusion of additional streams of evidence did not substantively change the overall certainty of evidence rating, while being fairly resource intensive and requiring specific methodological expertise. However, we applied the framework to only four test cases out of a universe of at least dozens of potential interventions across the CDC's 15 PHEPR Capabilities, making it difficult to generalize our experience. And, where the certainty of evidence was upgraded, it significantly impacted the recommendations. Also, the other evidence streams also informed and impacted the Evidence to Decision portion of the implementation guidance and practice recommendations section of the framework (see Figure 2). Additional application of the framework will provide a better evaluation of the additional value of being inclusive of different evidence.

5 | CONCLUSION

The NASEM committee's GRADE adaption for mixed-methods evidence provides a system for integrating and assessing diverse streams of evidence into a single confidence in the evidence rating. The methods were subject to initial testing in four reviews of interventions for public health emergency preparedness. We hope that over time, and as reviewers gain more experience of using a single certainty of evidence rating for a finding that the NASEM committee's GRADE adaption for mixed-methods evidence will further evolve with rigorous testing. The GRADE methodology is continually refined through the work of the GRADE working groups, one of which is actively developing methods for assessing the certainty of the body of evidence for complex health and social interventions.[38–42] Further testing is needed. Just as the GRADE system used in 2021 is a different and much-improved version of the original GRADE system proposed in 2001, we expect that over time further testing wll identify aspects of our proposed method that can be improved. The adaption has yet to be endorsed by the GRADE working group.

AUTHOR CONTRIBUTIONS

Conception and design of study: Ned Calonge, Paul G. Shekelle, Douglas K. Owens, Steven Teutsch, Autumn Downey, Lisa Brown, Jane Noyes. *Conduct of study and interpretation of data*: Ned Calonge, Paul G. Shekelle, Douglas K. Owens, Steven Teutsch, Autumn Downey, Lisa Brown, Jane Noyes. *Manuscript writing*: Ned Calonge, Paul G. Shekelle, Douglas K. Owens, Steven Teutsch, Autumn Downey, Lisa Brown, Jane Noyes. *Critical revision and final approval of manuscript*: Ned Calonge, Paul G. Shekelle, Douglas K. Owens, Steven Teutsch, Autumn Downey, Lisa Brown, Jane Noyes. *Accountable for all aspects of the work*: Ned Calonge, Paul G. Shekelle, Douglas K. Owens, Steven Teutsch, Autumn Downey, Lisa Brown, Jane Noyes.

ACKNOWLEDGMENTS

The authros are grateful to the committee on Evidence-Based Practices for Public Health Emergency Preparedness and Response, all advisors (in particular Professor Holger Schunemann), stakeholders and all those who attended dissemination and engagement activities.

FUNDING INFORMATION

Funded by the US Centers for Disease Control and Prevention. The NASEM committee functioned independently to develop the methods.

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest that could be perceived as prejudicing the impartiality of the research reported.

DATA AVAILABILITY STATEMENT

All data of relevance are presented in the manuscript or NASEM report.[19]

ORCID

Ned Calonge https://orcid.org/0000-0002-2653-6001
Jane Noyes https://orcid.org/0000-0003-4238-5984

REFERENCES

1. Petticrew M, Knai C, Thomas J, et al. Implications of a complexity perspective for systematic reviews and guideline development in health decision making. *BMJ Glob Health*. 2019;4:e000899. doi:10.1136/bmjgh-2018-000899
2. Guise J-M, Chang C, Butler M, Viswanathan M, Tugwell P. AHRQ series on complex intervention systematic reviews—paper 1: an introduction to a series of articles that provide guidance and tools for reviews of complex interventions. *J Clin Epidemiol*. 2017;90:6-10.
3. Briss PA, Zaza S, Pappaioanou M, et al. Developing an evidence-based guide to community preventive services—methods. *Am J Prev Med*. 2000;18(1S):35-43.
4. Noyes J, Booth A, Moore G, Flemming K, Tuncalp O, Shakibazadeh E. Synthesising quantitative and qualitative evidence to inform guidelines on complex interventions: clarifying the purposes, designs and outlining some methods. *BMJ Glob Health*. 2019;4(Suppl 1):e000893.
5. Petticrew M, Anderson L, Elder R, et al. Complex interventions and their implications for systematic reviews: a pragmatic approach. *J Clin Epidemiol*. 2013;66(11):1209-1214.
6. Waters E, Hall BJ, Armstrong R, Doyle J, Pettman TL, de Silva-Sanigorski A. Essential components of public health evidence reviews: capturing intervention complexity, implementation, economics and equity. *J Public Health*. 2011;33(3):462-465.
7. Pluye P, Hong QN. Combining the power of stories and the power of numbers: mixed methods research and mixed studies reviews. *Annu Rev Public Health*. 2014;35:29-45. doi:10.1146/annurev-publhealth-032013-182440
8. Savoia E, Agboola F, Biddinger PD. Use of after action reports (after action reports) to promote organizational and systems learning in emergency preparedness. *Int J Environ Res Public Health*. 2012;9(8):2949-2963.
9. Howick J, Glasziou P, Aronson JK. Evidence-based mechanistic reasoning. *J R Soc Med*. 2010;103(11):433-441.
10. Hill AB. The environment and disease: association or causation? *Proc R Soc Med*. 1965;58(5):295-300.
11. Howick J, Glasziou P, Aronson JK. The evolution of evidence hierarchies: what can Bradford Hill's "guidelines for causation" contribute? *J R Soc Med*. 2009;102(5):186-194.
12. Thomas J, Harden A. Methods for the thematic synthesis of qualitative research in systematic reviews. *BMC Med Res Methodol*. 2008;8:45.
13. Glenton C, Colvin CJ, Carlsen B, et al. Barriers and facilitators to the implementation of lay health worker programmes to improve access to maternal and child health: qualitative evidence synthesis. *Cochrane Database Syst Rev*. 2013;2013(10):CD010414.
14. Harden A, Thomas J, Cargo M, et al. Cochrane qualitative and implementation methods group guidance series-paper 5: methods for integrating qualitative and implementation evidence within intervention effectiveness reviews. *J Clin Epidemiol*. 2018;97:70-78.
15. WHO. *Communicating Risk in Public Health Emergencies: A WHO Guideline for Emergency Risk Communication (ERC) Policy and Practice*. World Health Organization; 2018.
16. Nelson C, Lurie N, Wasserman J, Zakowski S. Conceptualizing and defining public health emergency preparedness. *Am J Public Health*. 2007b;97(S1):S9-S11.
17. Hunter JC, Yang JE, Crawley AW, Biesiadecki L, Aragon TJ. Public health response systems in-action: learning from local health departments' experiences with acute and emergency incidents. *PLoS One*. 2013;8(11):e79457. doi:10.1371/journal.pone.0079457
18. Carbone EG, Thomas EV. Science as the basis of public health emergency preparedness and response practice: the slow but crucial evolution. *Am J Public Health*. 2018;108(S5):S383-S386.
19. National Academies of Sciences, Engineering, and Medicine. *Evidence-Based Practice for Public Health Emergency Preparedness and Response*. The National Academies Press; 2020. doi:10.17226/25650

20. Moher D, Schulz KF, Simera I, Altman DG. Guidance for developers of health research reporting guidelines. *PLoS Med*. 2010;7(2):e1000217.
21. The Grading of Recommendations Assessment, Development and evaluation (short GRADE) working group. https://www.gradeworkinggroup.org/
22. Community Preventative Services Task Force. https://www.thecommunityguide.org/
23. US Preventive Services Task Force. https://www.uspreventiveservicestaskforce.org/uspstf/
24. National Aeronautics and Space Administration Integrated Medical model. https://www.nasa.gov/audience/foreducators/mathandscience/research/Prob_IMM_detail.html
25. Clearing House for Labor Evaluation Research (CLEAR). https://clear.dol.gov/
26. What Works Clearing House. https://ies.ed.gov/ncee/wwc/
27. Evaluation of Genomic Applications in Practice and Prevention (EGAPP). https://www.cdc.gov/egappreviews/default.html
28. National Highway Traffic Safety Administration. https://www.nhtsa.gov/sites/nhtsa.gov/files/documents/812478_countermeasures-that-work-a-highway-safety-countermeasures-guide-.pdf
29. Cochrane risk of bias tool. https://epoc.cochrane.org/sites/epoc.cochrane.org/files/public/uploads/Resources-for-authors2017/suggested_risk_of_bias_criteria_for_epoc_reviews.pdf
30. Critical Appraisal Skills Programme (CASP). 2018. CASP appraisal checklists. http://casp-uk.net/casp-tools-checklists
31. Lewin S, Glenton C, Munthe-Kaas H, et al. Using qualitative evidence in decision making for health and social interventions: an approach to assess confidence in findings from qualitative evidence syntheses (GRADE-CERQual). *PLoS Med*. 2015;12(10):e1001895.
32. Tyndall, J. 2010. AACODS checklist. https://dspace.flinders.edu.au/xmlui/bitstream/handle/2328/3326/AACODS_Checklist.pdf
33. European Centre for Disease Prevention and Control (ECDC). *Best Practice Recommendations for Conducting after-Action Reviews to Enhance Public Health Preparedness*. European Centre for Disease Prevention and Control; 2018 https://www.ecdc.europa.eu/sites/default/files/documents/public-health-preparednessbest-practice-recommendations.pdf
34. National Transportation Safety Board. Aircraft Accident Report. https://www.ntsb.gov/investigations/AccidentReports/Reports/AAR0003.pdf
35. Barnett-Page E, Thomas J. Methods for the synthesis of qualitative research: a critical review. *BMC Med Res Methodol*. 2009;9(1):59.
36. Norris SL, Bero L. GRADE methods for guideline development: time to evolve. *Ann Intern Med*. 2016;165:810-811. doi:10.7326/M16-1254
37. The Community Guide. 2012. Emergency preparedness and response: school dismissals to reduce transmission of pandemic influenza: summary evidence tables—economic review. https://www.thecommunityguide.org/sites/default/files/assets/SET-schooldismissals-econ.pdf
38. Movsisyan A, Melendez-Torres GJ, Montgomery P. Users identified challenges in applying GRADE to complex interventions and suggested an extension to GRADE. *J Clin Epidemiol*. 2016;70:191-199.
39. Montgomery P, Movsisyan A, Grant SP, Macdonald G, Rehfuess EA. Considerations of complexity in rating certainty of evidence in systematic reviews: a primer on using the GRADE approach in global health. *BMJ Glob Health*. 2019;4(Suppl 1):e000848.
40. Schunemann HJ, Cuello C, Akl EA, et al. GRADE guidelines: how ROBINS-I and other tools to assess risk of bias in nonrandomized studies should be used to rate the certainty of a body of evidence. *J Clin Epidemiol*. 2018;111:105-114.
41. Norris SL, Rehfuess EA, Smith H, et al. Complex health interventions in complex systems: improving the process and methods for evidence-informed health decisions. *BMJ Glob Health*. 2019;4(Suppl 1):e000963.
42. Rehfuess EA, Akl EA. Current experience with applying the GRADE approach to public health interventions: an empirical study. *BMC Public Health*. 2013;13:9.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Calonge N, Shekelle PG, Owens DK, et al. A framework for synthesizing intervention evidence from multiple sources into a single certainty of evidence rating: Methodological developments from a US National Academies of Sciences, Engineering, and Medicine Committee. *Res Syn Meth*. 2022;1-16. doi:10.1002/jrsm.1582