Rating the quality of a body of evidence on the effectiveness of health and social interventions: a systematic review and mapping of evidence domains

Ani Movsisyan[a], Jane Dennis[b], Eva Rehfuess[c], Sean Grant[d] & Paul Montgomery[e]

[a]Department of Social Policy and Intervention, University of Oxford, Oxford OX1 2ER, UK

[b]London School of Hygiene & Tropical Medicine, London WC1E 7HT, UK

[c]Institute for Medical Informatics, Biometry and Epidemiology, Ludwig-Maximilians-University, Munich 81377, Germany

[d]RAND Corporation, Santa Monica, CA 90407-2138, USA

[e]Department of Social Policy, Sociology and Criminology, University of Birmingham, Edgbaston, Birmingham B15 2TT, UK

Corresponding author:
Ani Movsisyan
Centre for Evidence Based Intervention
University of Oxford
Barnett House
32 Wellington Square
OX1 2ER
Phone: +44 (0) 1865 270325
Email: ani.movsisyan@spi.ox.ac.uk

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1002/jrsm.1290

Abstract

Introduction. Rating the quality of a body of evidence is an increasingly common component of research syntheses on intervention effectiveness. This study sought to identify and examine existing systems for rating the quality of a body of evidence on the effectiveness of health and social interventions.

Methods. We used a multi-component search strategy to search for full-length reports of systems for rating the quality of a body of evidence on the effectiveness of health and social interventions published in English from 1995 onward. Two independent reviewers extracted data from each eligible system on the evidence domains included, as well as the development and dissemination processes for each system.

Results. Seventeen systems met our eligibility criteria. Across systems, we identified thirteen discrete evidence domains: study design, study execution, consistency, measures of precision, directness, publication bias, magnitude of effect, dose-response, plausible confounding, analogy, robustness, applicability and coherence. We found little reporting of rigorous procedures in the development and dissemination of evidence rating systems.

Conclusion. We identified 17 systems for rating the quality of a body of evidence on intervention effectiveness across health and social policy. Existing systems vary greatly in the domains they include and how they operationalise domains, and most have important limitations in their development and dissemination. The construct of the quality of the body of evidence was defined in a few systems largely extending the Grading of Recommendations Assessment, Development and Evaluation (GRADE) approach. GRADE was found to be unique in its comprehensive guidance, rigorous development and dissemination strategy.

Keywords: intervention effectiveness, systematic review, guideline, evidence rating, GRADE, public health

1. Introduction

Rating the quality of a body of evidence is an increasingly common component of systematic reviews and practice guidelines on intervention effectiveness. While assessing risks of bias in each *individual* study included in a research synthesis is an important and well-established practice,[1,2] rating the quality of a body of evidence is a comparatively new practice that indicates the credibility and trustworthiness of the *totality* of evidence across studies in relation to a specific research question.[3,4] Systems for rating the quality of a body of evidence have been predominantly discussed and applied in health-related systematic reviews and clinical guideline development.[5,6] The Cochrane Collaboration was the first organisation to attempt to integrate the rating of a body of evidence as a mandatory procedure in research syntheses on intervention effectiveness. Specifically, Cochrane mandated use of the Grading of Recommendations Assessment, Development and Evaluation (GRADE) approach in the conduct of Cochrane intervention reviews.[4] Over the last decade, GRADE and other approaches for rating the quality of a body of evidence have proliferated. The GRADE approach, specifically, is currently used by over 80 organisations worldwide.[7]

Systems for rating the quality of a body of evidence typically involve an examination of various characteristics of evidence that ultimately results in a rating of that body of evidence. For example, in the GRADE approach, the process of rating starts with a consideration of the designs of included studies: if the body of evidence contributing to an outcome consists of randomised controlled trials (RCTs), the quality of a body of evidence is initially given a rating of "high", while a body of evidence consisting of non-randomised studies (NRS) is initially given a rating of "low".[8] The body of evidence is then assessed by considering eight further domains. Assessments within five domains—risk of bias, indirectness, inconsistency, imprecision and publication bias—are used to downgrade the initial rating. For a body of evidence consisting of NRS, assessments within the three remaining domains—magnitude of the effect, dose-response relationship in the effect, and counteracting plausible residual bias or confounding—may be used to upgrade the initial "low" rating. Quality ("certainty" is also another frequently used term) of evidence is ultimately categorised into one of four ratings—high, moderate, low and very low—that reflect the extent to which the review authors are confident or certain that an estimate of the effect for a specific outcome is correct.[8]

As use of evidence rating systems has increased, so have reports of challenges faced by those attempting to use these systems—particularly for research syntheses on social and public health interventions, which are often described as "complex".[9-12] Interventions are viewed as complex for a variety of reasons. Some dimensions of complexity are ascribed to aspects of the interventions themselves,[13,14] such as interventions with multiple components that aim to address different and multiple causes of the problems (e.g., both biological and social). Other dimensions of complexity are seen as emanating from system properties;[15] that is to say, long, non-linear, and dynamic relationships between interventions and outcomes, interactions and interdependencies between different components of interventions, and levels of target.[16] Consideration of complexity may require additional guidance when rating the quality of a body of evidence.[11,12,17] Study design is often a key issue, given that RCTs are not feasible or appropriate for many population-level interventions. In addition, many

researchers acknowledge that multifaceted heterogeneity between studies in systematic reviews of complex interventions is a more difficult type of problem, and requires specific procedures of planning and analysis.[18] There are also concerns that narrow perspectives on evidence synthesis and the process of rating the quality of a body of evidence with simple hypotheses about the causal relationships may result in naïve and misleading synthesis results.[19,20] Furthermore, there are ambiguities around how best to conceptualise and interpret the construct of the quality of the body of evidence on the effectiveness of an intervention, when effects are contingent upon intervention programming, implementation and contextual factors.[12]

1.1. *Objectives*

In view of the challenges reported in applying quality of a body of evidence rating outside of biomedical settings and interventions,[11,12] this paper sets out to systematically review systems for rating the quality of a body of evidence on intervention effectiveness, including systems from health and social policy. Previous systematic reviews investigating evidence rating systems have mainly focused on scientific evidence in biomedical contexts and have not included systems from social policy domains such as public health, education, and crime and justice.[21,22]

The key objectives of this systematic review therefore are to:

1) Identify existing systems for rating the quality of a body of evidence on intervention effectiveness across health and social policy;

2) Examine how these systems describe the construct of the quality of a body of evidence and map out discrete domains they use to rate that quality;

3) Describe the reported procedures used to develop and disseminate the systems.

The resultant "state of the field" map of the systems can be used by any reviewer to identify and adopt systems and domains for rating the quality of a body of evidence that are relevant for their specific needs.

2. Methods

2.1. Eligibility criteria

Methods of this systematic review are described in detail in an *a priori* developed protocol (see Supplementary File 1). To be included in this review, a system had to (1) comprise a full-length document reporting a procedure for rating the quality of a *body of evidence*, derived from evidence synthesis that integrates results across individual studies on the effectiveness of health or social interventions; (2) be published in English from 1995 onward, when evidence rating was first proposed as a stage of research synthesis.[23] Where a document discussed a system developed by others (e.g., a literature review), we retrieved the original documents reporting those systems and examined them for eligibility. We excluded documents if they described a procedure for rating the quality of a body of evidence on intervention effectiveness for a specific clinical topic (e.g., systems used in specific guidelines on osteoarthritis, brain injury), as these are largely covered by the two previous systematic reviews.[21,22] We also excluded systems that were no longer used by an

organisation (e.g., the systems previously used by the Scottish Intercollegiate Guidelines Network and the Institute for Clinical Systems Improvement, before these organisations adopted the GRADE approach). Information on suspended use of these systems was either directly available on the organisation's website, or was obtained through email communication with representatives of the organisation.

2.2. Systematic search strategy

We used a multi-component search strategy with multiple sources in an attempt to maximise the sensitivity of the search. First, we updated search strategies used in previous systematic reviews[21,22] and expanded them to include social science databases. We ran these searches on 2 June 2016 in the following databases: Applied Social Sciences Index (ASSIA), Cochrane Methodology Register (Cochrane Library), EMBASE, MEDLINE, PsycINFO, SCIE Social Care Online, Scopus Social Sciences, and Social Sciences Citation Index (Web of Knowledge). Next, using the expertise of the authors and through bibliography searches of the related literature we located and searched the websites of 83 key stakeholder organisations that specifically aim to aggregate, review, and assess evidence across social policy domains, such as child and family welfare, international development, crime and justice, public health, and education (see Supplementary File 1 for the search strategy). Thirdly, we searched bibliographies of all the included documents and literature reviews containing secondary reporting of eligible systems. Finally, we consulted experts identified from the website searches to check whether we missed any systems.

Screening of all titles, abstracts and full-text documents was conducted by the first author (AM) using the Rayyan web application for systematic reviews.[24] A subset of randomly chosen titles (10%) was independently screened by a second author (JD). All discrepancies were discussed until agreement was reached.

2.3. Data extraction

We extracted data on four types of information. First, we extracted descriptive information about included systems, namely the author, year, title, publication source, and eligibility criteria. We then extracted information from each system on how its authors defined the construct of the quality of a body of evidence. We further extracted details of specific domains within the system used to rate the quality of a body of evidence, how these domains were defined, and how ratings of the quality of a body of evidence were categorised (e.g., "high", "moderate", or "low"). Extending the pre-specified domains for development and dissemination of research reporting guidelines,[25,26] we looked at whether systems reported any preparatory activities, such as a review of literature on existing domains for rating a body of evidence and consensus based activities, such as a Delphi exercise and expert meetings. Finally, we looked for information on how the documents describing the systems were written up and disseminated, such as whether the authors of the systems described how they planned to address criticism and feedback for the system or whether the system was available on an open-access website.

The first author (AM) and a second independent reviewer (either JD or a research assistant) extracted information about the content, development and dissemination of the

included systems into a Microsoft Excel spreadsheet. Three independent reviewers (AM, JD, ER) piloted the data extraction form on the same evidence rating system before continuing with the remaining systems.

2.4. Data synthesis

We employed a three-step procedure to describe the domains of evidence for rating the quality of a body of evidence in the included systems. First, we created an inventory of all identified domains by using cross-case tables.[27] We examined these tables to compare how the domains for rating the quality of a body of evidence were labeled, defined, and operationalised across included systems. We then compiled a discrete (i.e., non-redundant) list of domains of evidence considered in the included systems. Systems used different terminology to denote similar constructs and domains of evidence (for example, aspects of the domain that is termed as "imprecision" in the GRADE approach were covered by "precision" in the system used by the Agency for Healthcare Research and Quality (AHRQ) and fell under the domain termed "clinical impact" in the system adopted by the National Health and Medical Research Council (NHMRC) of Australia). Where such overlap existed, we mainly followed the terminology of the GRADE approach to describe the discrete set of domains. We supplemented this with a list of additional domains that are not currently considered in the GRADE approach, but were found in other systems, and followed the terminology used in the systems to describe these domains. Finally, to help readers to visualise findings, we created a heat map summarising how the systems reported the identified discrete domains of evidence (see Figure 1). By using different colour shades, the heat map describes whether these domains of evidence are reported in each included system or not. Where a system reported the domain and yet did not provide specific criteria and guidance for rating it, the map denotes those as a different category of reporting (i.e., with a brighter shade). Similar to this, we developed a second heat map describing how authors reported activities underpinning the development and dissemination of the included systems (see Figure 2). Both of these heat maps were developed by one author (AM) and further verified by the second author (JD).

3. Results

We identified 11,758 records after duplicates were removed. After title and abstract screening, we assessed the full-texts of 141 records, from which 28 records were found to be eligible for inclusion in this review. Overall, these 28 records describe 17 evidence rating systems (see Figure 3 for the PRISMA flow diagram).

3.1. Excluded studies

Of the 113 records excluded at full text, 45 involved literature reviews of evidence rating systems, 28 were editorials or conference abstracts, and four records were not published in English (Chinese, French, Portuguese and Spanish). Twenty-nine records described procedures and domains for categorising interventions on websites of different "what works" organisations, also known as evidence clearinghouses or evidence-based programme registers.[28] Because these procedures and corresponding domains of evidence did

not consider a "body of evidence", we excluded them from this review (a full list of these systems and their specific domains is available from the first author upon request). Through website searches and contacts with experts we established that six systems were no longer used.[23,29-33] A further system, the Confidence in the Evidence from Reviews of Qualitative Research (CERQual), which is designed for sole application to a body of qualitative evidence, was not eligible for use in assessment of effectiveness evidence.[34]

3.2. Characteristics of the sample

Fourteen of the included systems were developed for healthcare, including general clinical and public health interventions (see Table 1). Only three systems were developed for other policy domains—specifically education, criminology, and international development.[35-37] Three of the included systems were largely based on the GRADE approach,[38] but introduced modifications that warrant their classification as separate systems.[39-41] Ten systems mentioned specific research synthesis methods for which the system was developed; most referred to a meta-analysis or a "narrative synthesis" without a single pooled effect estimate[42] to synthesise data on the effects of an intervention. Only one system was explicitly described for use with a mixed-method approach to research synthesis.[36] Eight of the systems described rating the quality of a body of evidence primarily within the context of research syntheses only, while eight others described rating the quality of a body of evidence for a guideline development context. Only the GRADE approach addressed the conceptual and procedural differences when using the domains of evidence for assessing a body of evidence for research synthesis versus guideline development contexts.[38]

We identified inconsistencies in how included systems labelled and defined rating of the quality of a body of evidence overall and the components of that rating. The most frequently used terms to describe the overall rating of the quality of a body of evidence were *strength of evidence, grades of evidence, quality, confidence,* or *certainty in evidence.*[37,38,40,43-47] In contrast, the most commonly used terms for assessing the conduct of individual included studies were *levels of evidence, critical appraisal, quality appraisal, study limitations, risk of bias,* and *study quality*.[37,43,44,48] From these, terms such as *levels of evidence, risk of bias* and *study limitations* were mainly discussed with regard to assessing studies for bias and internal validity, while *study quality, quality appraisal and critical appraisal* were used to denote study execution more broadly with regard to eliminating threats to both internal and external validities.

3.3. Defining quality of a body of evidence

Only six systems—three of which are largely based on the GRADE approach—provided a definition for the construct of the quality of a body of evidence on intervention effectiveness.[38-41,46,47] In a systematic review context, the GRADE approach and three derivative systems defined quality of a body of evidence as *"the extent of confidence that an estimate of the effect is correct"*.[38-40] The Guide to Community Preventive Services defined quality of a body of evidence as *"confidence that changes in outcomes are attributable to the interventions"*,[46] and the U.S. Preventive Services Task Force (USPSTF) as the *"likelihood that the assessment of the net benefit (i.e., benefits minus harms) of a preventive service is correct"*.[47] The USPSTF definition is similar to how the GRADE approach defines the

overall quality of a body of evidence in the context of guideline development, when considering all important outcomes associated with the intervention, including harms. In this context GRADE defines the overall quality of a body of evidence as *"the extent of confidence that an estimate of the effect is adequate to support a particular decision or recommendation"*.[49]

In order to assess the net benefit of a preventive service, the USPSTF system uses analytic frameworks, also called "chain of evidence" diagrams to map out the specific linkages in the overall chain of evidence that must be present for a preventive service to be considered effective.[47] The system assesses the quality of a body of evidence for each separate linkage in the chain of evidence to draw conclusions about the overall effectiveness of a preventive service. This approach is very similar to that adopted by the GRADE-modified Grading of Evidence for Public Health Interventions (GEPHI) system.[41] In addition to rating the quality of a body of evidence for the estimates of the effect of an intervention (which corresponds to the approach described in GRADE), the GEPHI system suggests to also rate the quality for the overall causal chain of an intervention. This rating of the confidence in the overall causal chain of an intervention is referred to as *coherence of evidence* assessment in the GEPHI system.[41]

3.4. Mapping of evidence domains

The evidence domains used to rate the quality of a body of evidence were often similar in concept across systems yet different in how they were described and operationalised. We encourage readers to use Table 1 and Figure 1 as two complementary sources of information on the identified evidence grading systems to examine the discrepancies in labelling and describing evidence domains. Table 1 provides an overview of the domains of evidence as they are reported in the original studies, while Figure 1 maps the thirteen discrete domains we identified in included systems and presents how they are reported in each of the included systems. More information on how the specific evidence domains were defined and operationalised in each system is presented in Supplementary File 2. In the sections below, we briefly summarise the identified discrete set of domains of evidence (see Figure 1), as well as the reported activities underpinning the development and dissemination of these systems (see Figure 2).

3.4.1. Study design

Twelve systems included an evidence domain related to the design of the individual studies constituting the body of evidence. All but four of these systems[35,36,45,50] described an "evidence hierarchy" approach that influenced how overall quality of a body of evidence was assessed. Procedurally, this entailed initially privileging a body of evidence from certain study designs (namely RCTs) as providing a higher quality (compared to other study designs) before assessing other evidence domains. While all systems with an evidence hierarchy approach placed evidence from RCTs at the top of this hierarchy, many further privileged specific non-randomised study designs over others. For example, the system used by the Joanna Briggs Institute[39] suggested initial ratings of quality depending on whether a body of evidence consists of experimental (Level 1), quasi-experimental (Level 2), or observational

studies (Level 3). Similarly, the GRADE-modified GEPHI system for public health interventions recommends that a body of evidence consisting of non-randomised studies with controls or before and after [uncontrolled] studies have an initial rating of "moderate" quality if these studies used methods to minimise selection bias and confounding.[41]

3.4.2. Study execution

Fifteen systems included an evidence domain related to assessing how well studies constituting the body of evidence were executed to minimise threats to internal and external validities (also labelled as *quality of study execution*, *risk of bias*, *study limitations*, *study quality*). In the majority of instances, however systems mainly included criteria to assess risks of bias or threats to the internal validity for assessing study execution. A few systems, however also included specific criteria for assessing the generalisability of the study results, that is, criteria related to the external validity of the individual studies in the body of evidence.

Systems varied in how they operationalised assessment of study execution. Some systems used design-specific criteria, such as checklists or signaling questions for appraising RCTs[36,38,40,43] or longitudinal studies.[43] Most systems, however, described more generic criteria to assess study execution across various study designs included in the body of evidence.[37,45,46,48,50]

3.4.3. Consistency

Fourteen systems included an evidence domain related to the *consistency* of evidence. Generally, systems defined consistency as "*the extent to which findings are similar across included studies"* in a body of evidence,[48] usually in reference to the degree of similarity in the magnitude and/or direction of effect estimates. The majority of the systems, however did not report any specific criteria on how to rate consistency in the body of evidence. Only a few systems discussed specific procedures, such as statistical testing for heterogeneity to rate consistency in the body of evidence. The GRADE-modified GEPHI approach distinguished between two types of consistency ratings:[41] the first type was identical to the domain of the GRADE approach termed as *inconsistency* and defined as *"assessment of statistical heterogeneity in the estimates of the effect"*.[51] The second type of consistency rating was specified in the system as "consistency" assessment and was defined as presence of *"consistent evidence across a large number of settings, geographical locations and diverse epidemiological study designs"*. The system argued that the fact that an intervention effect is reproducible under highly variable conditions suggests reduced likelihood that the observed effect is attributable to confounding or bias.[41] This can increase a reviewer's confidence in the body of evidence regarding the overall effectiveness of an intervention.

3.4.4. Measures of precision

Eleven systems included an evidence domain that we have classified as relating to *measures of precision* of the body of evidence: i.e., considerations of the impact that random error may have on effect estimates. Systems differed widely in the level of specification and sophistication they required for assessing precision of the body of evidence. For instance, many systems recommend only considering the number of studies in the body of evidence as

a measure of precision,[37,43,45,46,52] however only one of these systems specifies a threshold for the minimum number of studies to be included in the body of evidence.[52] Furthermore, only the GRADE approach and its variants described specific criteria for assessing precision with regard to the sufficiency of the sample size of the body of evidence.[38-41] These systems assessed sufficiency of the sample size relative to an "optimal information size": i.e., *"number of patients (for continuous outcomes) and events (for dichotomous outcomes) that would be needed to regard a body of evidence as having adequate power"*.[53] In addition, these systems also considered the boundaries of confidence intervals for an effect estimate in relation to a null effect and a clinically important effect threshold to make an overall judgement about the precision of a body of evidence. The estimate of the effect of an intervention is judged to be less precise if the confidence interval is wide to include a null effect or a threshold, which is considered as clinically unimportant.[53]

3.4.5. Directness

In general, systems used concepts of *directness*, *applicability* and *generalisability* of evidence interchangeably and inconsistently—often without providing clear definitions or specific criteria to guide the assessment.[35,37,47,48,50] In addition, these terms were not necessarily used as synonyms across systems. For example, the system endorsed by the National Health and Medical Research Council (NHMRC) of Australia used the term "applicability" to address whether the body of evidence was relevant to the local context (including the organisational and cultural context), while the term generalisability was used to refer to how precisely a body of evidence answered a review or a guideline question in terms of populations and settings of interest.[48] In order to disentangle the discrepancies in the terminology, we have used the terminology of the GRADE approach, namely "directness" of evidence to describe the domains of evidence from the included systems related to the notion of comparability of the evidence to the original research question. We have identified six systems that used this domain of evidence to assess how directly the available evidence answers a review or a guideline question with regard to the Population, Intervention, Comparison and Outcomes (PICO) elements of the question.[35,39-41,48,54]

3.4.6. Publication bias

Five systems included *publication bias* as a domain for rating the quality of a body of evidence.[36,39-41,55] All but one of these systems followed a definition of publication bias as used within the GRADE approach, that is *"a failure to identify studies as a result of studies remaining unpublished or obscurely published"*.[55] The system used by AHRQ on the other hand, considered publication bias as only one type of potential bias within a broader domain of reporting biases, which was itself defined as a decision by authors or journals to report research findings based on their direction and magnitude of effect.[40] Selective outcome reporting and selective analysis reporting were the other types of reporting biases described in this system.

3.4.7. Magnitude of effect

We identified seven systems, which included *magnitude of effect* as a distinct domain to rate the quality of a body of evidence on the effectiveness of health or social

interventions.[36,39-41,43,46,56,57] However, only four of these systems specified the thresholds for what they considered to be a "large" magnitude of effect.[39,41,56,57] This predominantly included a relative risk (RR) greaten than 2, or less than 0.2, as suggested in the GRADE approach.[56]

3.4.8. Dose-response

Overall, five systems considered *dose-response* as a distinct domain of evidence when rating the quality of a body of evidence on the effectiveness of health or social interventions.[39-41,47,56,57] Systems commonly defined dose-response as a *"pattern of a larger effect with greater exposure to an intervention"*.[40]

3.4.9. Plausible residuals

All systems that followed the structure of the GRADE approach (overall four systems, including GRADE itself) considered counteracting confounding, as a domain to upgrade the quality of a body of evidence, when a body of evidence is mainly comprised of observational studies.[39-41,56] Two possibilities were commonly applied: *"if all plausible residual biases would diminish the observed effect, or if all plausible residual biases would suggest a spurious effect when no effect is observed"*.[56]

3.4.10. Analogy

Only one system—the GEPHI system—included an evidence domain related to analogous evidence. The GEPHI system operationalised analogous evidence as supporting evidence from similar or "analogous" interventions that are known to operate through the same or similar mechanisms, which if present could lead to a higher quality of a body of evidence rating.[41] In the context of WHO guidelines on indoor air quality, the system discusses the example of how certainty in the effects of household air pollution from solid fuel can be enhanced by strong empirical evidence about the effects of second-hand or active smoking. In this example, both household air pollution and second-hand or active smoking expose individuals to similar combustion mixtures, and therefore are viewed as analogous pieces of evidence.[41]

3.4.11. Robustness

Robustness of evidence was described as a domain to rate the quality of a body of evidence by one system.[52] The system suggests that reviewers measure robustness of evidence through sensitivity analysis with a priori defined thresholds. For example, a reviewer may decide a priori that a threshold for robustness assessment is one in which *"confidence intervals of the last three cumulative, random-effects meta-analyses remain fully on the same side of zero after removing of the study with the smallest weight"*.[52]

3.4.12. Applicability

Four systems described applicability as a domain of evidence measuring the extent to which evidence may be applicable in a specific context.[37,47,48,50] It is worth highlighting that we identified three additional systems,[40,45,46] which considered applicability of evidence as a separate judgement when making recommendations for practice. In these systems, discussion of applicability was held separately from other domains of evidence, and largely within a

context of guideline development. For example, the GRADE-based system endorsed by AHRQ clearly separates judgments of directness of evidence from that of applicability assessment. In this system, directness of evidence is defined to express *"how closely the available measures an outcome of interest"*, and relies on two judgments:[40] the directness of the employed outcomes (i.e., whether the available evidence is in fact only a proxy for an ultimate outcomes of interest) and directness of comparisons (i.e., whether evidence derives from head-to-head comparisons). Meanwhile, the system defines applicability as the external validity of the evidence base with regard to different populations, and is considered explicitly, but separately from the overall rating of the quality of a body of evidence.[40]

3.4.13. Coherence

Only three systems included an evidence domain related to assessing the coherence of the causal pathway of an intervention:[41,47,57] that is, related to the assessment of a theory of change or a mechanism whereby an intervention is expected to operate. The GEPHI system recommends assessing confidence in the overall causal pathway between an intervention and distal outcomes (referred to as rating of *coherence* of evidence) with regard to the evidence informing each individual link in the causal pathway.[41] It describes this domain specifically in the context of interventions that involve complex causal pathways, where evidence directly linking the intervention with the distal outcomes is frequently unavailable. Similarly, by using analytic frameworks the USPTSF system rates certainty of evidence in the overall chain of evidence for a specific preventive service.[47] The system described by Tang and colleagues (2007) included assessment of the known mechanisms of action as a domain of evidence for rating of the quality of a body of evidence: "*if the theoretical basis is not known, the strength of evidence will be less convincing"*.[57]

3.5 Development and dissemination of the evidence rating systems

Figure 2 describes how authors report procedures underpinning the development and dissemination of the systems. With regard to the preparatory activities for developing the system, only four systems empirically demonstrated the need for developing a new evidence rating system by referring to a separate publication by the same research team providing a critical appraisal of existing systems.[38,43,45,48] More frequently systems reported participants involved in the development of the system, and only four systems described obtaining funding for developing the system.[36,48,50,52] None reported conducting a Delphi process to develop the system, and only five reported hosting an expert meeting. However, with the exception of the GRADE approach, these systems did not provide further details on how these meetings were organised.[38,44,48,50,58] The GRADE Working Group, on the other hand, organises annual meetings lasting two to three days, where members of the group have an opportunity to meet face-to-face and further discuss and develop and refine aspects of the GRADE methodology.[7]

Regarding the write-up and dissemination activities, only three systems described how the publication introducing the system was developed,[44,48,50] while instructions for using the systems were predominantly described in the same document that introduced it. In six instances, willingness to incorporate the feedback of users and update the systems was mentioned.[37,38,40,43,44,48] Finally, although the majority of the systems are available online,

information regarding adherence to or translation of the systems—was not reported for any system except for GRADE (further details on this can be found on the website of the GRADE Working Group).[7] The GRADE approach was also unique in involving ongoing working groups aiming to continually advance and expand the applicability of its methodology in step with developments in the area of evidence synthesis and assessment.

4. Discussion

4.1. "State of the field" map of evidence rating systems for health and social interventions

This systematic review set out to describe the content, development and dissemination of the systems for rating the quality of a body of evidence on intervention effectiveness across health and social policy. The review identified 17 systems that have made useful contributions to rating the quality of a body of evidence in health and social research synthesis. While this review identified domains of evidence that were commonly reported across the systems, there was significant variation in the specifications for these domains. Systems used different terminology to denote similar constructs of evidence when rating the quality of a body of evidence. Systems also varied in how they operationalised the domains of evidence, that is, in whether they described specific criteria and provided guidance for assessing each domain in an operationalisable manner. This review also identified domains of evidence that were found only in a few systems (see Figure 1). In general, the discrete set of domains identified in our review can be viewed to largely follow the "viewpoints for causation" proposed by Sir Austin Bradford Hill,[59] although the relative coverage of these criteria across the included systems varies. For example, domains of evidence that will correspond to the Hill's criteria of experiment (study design and study execution), strength of association (magnitude of effect), consistency and dose-response gradient have been reported more extensively in evidence rating systems. Meanwhile, our review found only three systems, which considered domains corresponding to the Bradford Hill viewpoints of plausibility and coherence of evidence, and only one system included a domain on the analogous evidence. This can partly be explained by the challenges of developing an operational framework in research synthesis to assess the evidence against these criteria, including the need to search and integrate different sources of evidence.[60]

As this systematic review aimed to consider evidence rating systems across health and social policy, the identified variation in the terminology and description of evidence domains may partly reflect how research synthesis and its practice differs across policy areas and types of interventions. One of the most contested topics in the discussions of the quality of a body of evidence relates to the hierarchy of evidence initially described in the paradigm of evidence-based medicine as an approach to differentiate between weak and strong study designs for assessing intervention effectiveness.[61] While different versions of the evidence hierarchy have been described in clinical medicine, all of them place study designs such as case series (considered relatively weaker in terms of protecting against threats to internal validity) in the bottom of the hierarchy, followed by case-control and cohort studies in the middle and RCTs at the top.[62] As our findings demonstrate, this evidence hierarchy approach is still used in many evidence rating systems, and particularly those developed and employed in clinical medicine. The widely adopted GRADE approach also follows this approach by

way of describing two broad categories of study designs as a starting point for the body-of-evidence rating process (RCT evidence is initially rated as "high" quality and non-RCT evidence as "low" quality). By contrast, our findings show that systems which are used in broader policy areas, such as public health, tend to allow more flexibility for differentiating between the many types of non-RCT designs within their constructions of evidence hierarchies (see section 3.4.1 and Table 1). This practice is commensurate with a view that quasi-experimental approaches should be given appropriate provisions in evidence rating systems as valuable methods for making causal inferences for public health interventions.[63]

Consistency of the body of evidence was another frequently reported domain of evidence in the included systems. Our findings demonstrate that evidence rating systems currently conceptualise consistency as similarity in the magnitude and direction of effect estimates across studies (of same or similar design) included in the body of evidence. There are however concerns that this approach only partly reflects the central tenet of scientific method, specifically that findings are replicable across "a variety of situations and techniques".[59] From this perspective, there are suggestions for a broader interpretation of the consistency of evidence to also consider "triangulation of evidence" across different methodological approaches when arriving at overall conclusions about intervention effectiveness.[64] Triangulation has been defined as integration of evidence from several different methodological approaches (different study designs and analytical approaches), which address the same underlying causal question, but which vary in terms of key sources of potential bias (for example, multivariable regression, instrumental variables and RCTs).[65] The importance of evidence triangulation has been cogently argued in the context of public health interventions involving longer causal pathways and multiple targets and behaviours, such as smoking or alcohol consumption, which are difficult (or impossible) to evaluate with RCTs alone. When the results from different methodological approaches are consistent in that they all point to the same conclusion, this is argued to strengthen the confidence in the overall findings (see Lawlor et al., 2016).[65] Our review identified only one system which extended the domain of consistency to consider evidence from different study designs.[41] Its broad interpretation, which looks at evidence from different methodological approaches to inform the rating of the quality of a body of evidence was unique within our findings (see section 3.4.3).

Our review identified very few instances where systems provided a definition for the construct of the quality of the body of evidence (see section 3.3). The few reported definitions mainly focus on the confidence in a direct estimate of the effect of an intervention—a definition initially suggested by GRADE. It is worth noting here, that the most recent publication of the GRADE Working Group clarifies this definition of the quality of a body of evidence based on a priori defined threshold and the context of the review.[66]The quality of a body of evidence is currently conceptualised to reflect the extent to which reviewers can be confident that "the true effect for a specific outcome lies on one side of a specified threshold or within a chosen range".[66] The revised guidance suggests three types of ratings: non-contextualised, partly contextualised and fully contextualised (see Table 2 for more details). In this new conceptualisation, the quality of a body of evidence ratings are explicitly acknowledged to be contingent upon a priori defined thresholds of what may be

considered as meaningful effects in different contexts. These thresholds and the resultant ratings may therefore vary depending on the context and purpose of the review.

With regard to the activities underpinning the development and dissemination of the included systems, our review found that the majority of the systems did not report a comprehensive literature review or a consensus-based procedure for developing the system (see Figure 2). In a similar vein, we found little reporting of how these systems were written up and further disseminated. It therefore remains difficult to assess how the described domains of evidence have been conceptualised and the degree to which they are, or are not, the product of scientific consensus. In the meantime, if not properly developed and disseminated, these systems may have limited value and use in research synthesis.[26] In this regard, our review shows that the GRADE approach is one of the most comprehensive and transparent evidence rating systems in its guidance as well as its development and dissemination.[7]

4.2. Strengths and limitations

This review's unique contribution may lie in its thorough exploration of the content, development and dissemination of the existing systems for rating the quality of a body of evidence across a range of policy areas, following systematic searches of bibliographic databases and sources of grey literature. Consequently, this review provides a comprehensive inventory of evidence domains considered when assessing quality of a body of evidence in research syntheses on intervention effectiveness across not just health, but social policy as well. Considering the acknowledged challenges associated with locating evidence rating systems through formal literature searches,[22] we decided to balance the searches of scientific databases with an extensive search of grey literature, including 83 websites and databases of key stakeholder organisations. Furthermore, we complemented these searches with expert consultations to help locate these additional sources.

We note several limitations worth considering when interpreting our findings. First, we had to limit the scope of our review because of practical considerations. For instance, we included documents published in English only, and therefore might have missed relevant work from the non-English literature. Furthermore, given the identified variation in the terminology of the evidence domains, the mapping of these domains necessarily involved a degree of interpretation. It is therefore possible that another team of reviewers might have produced a different mapping of the domains with different conceptual categories. For example, another review team may have interpreted the broad evidence domain of the "efficacy data" of the Highest Attainable Standard of Evidence (HASTE) system[58] as referring to the strength of association and therefore in the map classified under the category of the "measures of precision", rather than consistency as we currently did. To address this concern, the initial mapping of evidence domains by the first author was independently verified by a second reviewer, and all issues were further discussed and clarified in the team.

4.3. Concluding remarks

The mapping of evidence domains presented in this review aims to clarify how domains of evidence for rating the quality of a body of evidence on intervention effectiveness have been specified, developed and disseminated across health and social policy. We see two broad applications of our mapping of evidence domains. First, it can serve as an aid for researchers to help choose the evidence rating system and corresponding domains of evidence most suitable for their research focus and context of work. Second, by delineating important gaps in the content, development and dissemination of current systems, it can indicate areas that may need further methodological development. It is worth noting that our mapping of domains should not be regarded as an expert advice on the best system for assessing the quality of a body of evidence on intervention effectiveness, but rather should be considered as a "state of the field" description and interpretation of the content and the processes of development and dissemination based on the information reported in the included systems.

Acknowledgements

We thank Agnes Ebenberger for her assistance with data extraction. This project was prepared as part of the GRADE Guidance for Complex Interventions, funded by the Economic and Social Research Council (ES/N012267/1). All the authors are current members of the GRADE Working Group. ER is a co-author of one of the evidence rating systems analysed as part of this work. SG's spouse is a salaried-employee of Eli Lilly and Company, and owns stock. SG has accompanied his spouse on company-sponsored travel.

References

1. Higgins JPT, Green S. (2011) Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0. Retrieved from: http://www.handbook.cochrane.org/. Accessed May 23, 2017.

2. Juni P, Altman DG, Egger M. Systematic reviews in health care: Assessing the quality of controlled clinical trials. *BMJ.* 2001;323(7303):42-46.

3. 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):924-926.

4. Higgins J, Lasserson T, Chandler J, Tovey D, Churchill R. *Methodological expectations of Cochrane interventions reviews.* London: Cochrane; 2016.

5. Gough D, Oliver S, Thomas J. *An introduction to systematic reviews.* London, UK: SAGE Publications Ltd; 2012.

6. Sackett DL. *Evidence-based medicine: how to practive and teach EBM.* Edinburgh: Churchill Livingstone; 2000.

7. The Grading of Recommendations Assessment, Development and Evaluation (GRADE) Working Group. Retrived from http://gradeworkinggroup.org/. Accessed November 17, 2017.

8. Balshem H, Helfand M, Schunemann HJ, et al. GRADE guidelines: 3. Rating the quality of evidence. *J Clin Epidemiol.* 2011;64(4):401-406.

9. Barbui C, Dua T, van Ommeren M, et al. Challenges in developing evidence-based recommendations using the GRADE approach: the case of mental, neurological, and substance use disorders. *PLoS Med.* 2010;7(8).

10. Harder T, Abu Sin M, Bosch-Capblanch X, et al. Towards a framework for evaluating and grading evidence in public health. *Health Policy.* 2015;119(6):732-736.

11. 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.

12. Rehfuess EA, Akl EA. Current experience with applying the GRADE approach to public health interventions: an empirical study. *BMC Public Health.* 2013;13:9.

13. Craig P, Dieppe P, Macintyre S, Michie S, Nazareth I, Petticrew P. Developing and evaluating complex interventions: New guidance. Medical Research Council; 2008.

14. Lewin S, Hendry M, Chandler J, et al. Assessing the complexity of interventions within systematic reviews: development, content and use of a new tool (iCAT_SR). *BMC Med Res Methodol.* 2017;17(1):76.

15. Hawe P, Shiell A, Riley T. Theorising interventions as events in systems. *Am J Community Psychol.* 2009;43(3-4):267-276.

16. Diez Roux AV. Complex systems thinking and current impasses in health disparities research. *Am J Public Health.* 2011;101(9):1627-1634.

17. Murad MH, Almasri J, Alsawas M, Farah W. Grading the quality of evidence in complex interventions: a guide for evidence-based practitioners. *Evid Based Med.* 2016;22(1):20-22.

18. Pigott T, Shepperd S. Identifying, documenting, and examining heterogeneity in systematic reviews of complex interventions. *J Clin Epidemiol.* 2013;66(11):1244-1250.

19. Petticrew M. Time to rethink the systematic review catechism? Moving from 'what works' to 'what happens'. *Syst Rev.* 2015;4(36).

20. Petticrew M, Shemilt I, Lorenc T, et al. Alcohol advertising and public health: systems perspectives versus narrow perspectives. *J Epidemiol Community Health.* 2017;71(3):308-312.

21. Bai A, Shukla V, Bak G, Wells G. *Quality Assessment Tools Project Report.* Ottawa: Canadian Agency for Drugs and Technologies in Health; 2012.

22. West S, King V, Carey Tea. *Systems to rate the strength of scientific evidence. Evidence Report/Technology Assessment No. 47(Prepared by the Research Triangle Institute–University of North Carolina Evidence-based Practice Center under Contract No. 290-97-0011).* Rockville, MD: Agency for Healthcare Research and Quality: AHRQ Publication No. 02-E016; 2002.

23. Guyatt G, H., Sackett D, L., Sinclair J, C., Hayward R, Cook D, J., Cook R, J. Users' guides to the medical literature. IX. A method for grading health care recommendations. Evidence-Based Medicine Working Group. *JAMA.* 1995;274(22):1800-1804.

24. Ouzzani M, Hammady H, Fedorowicz Z, Elmagarmid A. Rayyan-a web and mobile app for systematic reviews. *Syst Rev.* 2016;5(1):210.

25. Grant SP, Mayo-Wilson E, Melendez-Torres GJ, Montgomery P. Reporting quality of social and psychological intervention trials: a systematic review of reporting guidelines and trial publications. *PLoS One.* 2013;8(5):e65442.

26. Moher D, Schulz KF, Simera I, Altman DG. Guidance for developers of health research reporting guidelines. *PLoS Med.* 2010;7(2):e1000217.

27. Miles BM, Huberman AM. *Qualitative data analysis: an expanded sourcebook.* Thousand Oaks, CA: Sage; 1994.

28. Burkhardt J, T., Schroter D, C., Magura S, Means S, N., Coryn C, L., S. An overview of evidence-based program registers (EBPRs) for behavioral health. *Eval Program Plann.* 2015;48:92-99.

29. Eccles M, Freemantle N, Mason J. North of England evidence based guidelines development project: methods of developing guidelines for efficient drug use in primary care. *BMJ.* 1998;316(7139):1232-1235.

30. Greer N, Mosser G, Logan G, Halaas GW. A practical approach to evidence grading. *Jt Comm J Qual Improv.* 2000;26(12):700-712.

31. Harbour R, Miller J. A new system for grading recommendations in evidence based guidelines. *BMJ.* 2001;323(7308):334-336.

32. Liddle J, Williamson M, Irwig L. *Method for evaluating research and guideline evidence.* Sydney, Australia: NSW Health Department;1996.

33. Weightman A, Ellis S, Cullum A, Sander L, Turley R. *Grading evidence and recommendations for public health interventions: developing and piloting a framework.* Health Development Agency; Support Unit for Research Evidence (SURE), Information Services, Cardiff University; 2005.

34. 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).

35. Gough D. Weight of evidence: a framework for the appraisal of the quality and relevance of evidence In: Furlong J, Oancea A, eds. *Applied and Practice-based Research.* Vol 22: Research Papers in Education; 2007:213-228.

36. Johnson S, D., Tilley N, Bowers K, J. Introducing EMMIE: an evidence rating scale to encourage mixed-method crime prevention synthesis reviews. *J Exp Criminol.* 2015;11:459-473.

37. DFID. *How to Note: Assessing the strength of evidence.* Department for International Development (DFID); 2014.

38. Guyatt G, Oxman A, D., Akl E, A., et al. GRADE guidelines: 1. Introduction-GRADE evidence profiles and summary of findings tables. *J Clin Epidemiol.* 2011;64(4):383-394.

39. Joanna Briggs Institute. *Supporting document for the Joanna Briggs Institute levels of evidence and Grades of Recommendations.* (2014). The Joanna Briggs Institute Levels of Evidence and Grades of Recommendations Working Party. Retrieved from: https://joannabriggs.org/assets/docs/approach/Levels-of-Evidence-SupportingDocuments-v2.pdf. Accessed October 26, 2016.

40. Berkman ND, Lohr K, Ansari M, et al. *Grading the strength of a body of evidence when assessing health care interventions for the effective health care program of the*

Agency for Healthcare Research and Quality: An update. Rockville, MD: Agency for Healthcare Research and Quality. AHRQ Publication No. 13(14)-EHC130-EF;2013.

41. Bruce N, Pruss-Ustun A, Pope D, Heather A, Rehfuess E. *WHO indoor air quality guidelines: household fuel combustion: Methods used for evidence assessment.* World Health Organization (WHO) guidelines: household fuel combustion-technical paper on evidence review methods; 2014.

42. Popay J, Roberts H, Sowden A, et al. *Guidance on the Conduct of Narrative Synthesis in Systematic Reviews: A Product from the ESRC Methods Programme.* Lancaster, UK; 2006.

43. Clark E, Burkett K, Stanko-Lopp D. Let Evidence Guide Every New Decision (LEGEND): an evidence evaluation system for point-of-care clinicians and guideline development teams. *J Eval Clin Pract.* 2009;15(6):1054-1060.

44. Ebell MH, Siwek J, Weiss BD, et al. Strength of recommendation taxonomy (SORT): a patient-centered approach to grading evidence in the medical literature. *Am Fam Physician.* 2004;69(3):548-556.

45. NICE. *Methods for the development of NICE public health guidance: Process and methods guides*: National Institute for Health and Care Excellence (NICE); 2012.

46. Briss PA, Zaza S, Pappaioanou M, et al. Developing an evidence-based Guide to Community Preventive Services-methods. The Task Force on Community Preventive Services. *Am J Prev Med.* 2000;18(1 Suppl):35-43.

47. Sawaya GF, Guirguis-Blake J, LeFevre M, Harris R, Petitti D, Force USPST. Update on the methods of the U.S. Preventive Services Task Force: estimating certainty and magnitude of net benefit. *Ann Intern Med.* 2007;147(12):871-875.

48. Hillier S, Grimmer-Somers K, Merlin T, et al. FORM: an Australian method for formulating and grading recommendations in evidence-based clinical guidelines. *BMC Med Res Methodol.* 2011;11:23.

49. Guyatt G, Oxman A, D., Sultan S, et al. GRADE guidelines: 11. Making an overall rating of confidence in effect estimates for a single outcome and for all outcomes. *J Clin Epidemiol.* 2013;66(2):151-157.

50. Turner-Stokes L, Harding R, Sergeant J, Lupton C, McPherson K. Generating the evidence base for the National Service Framework for Long Term Conditions: a new research typology. *Clin Med (Lond).* 2006;6(1):91-97.

51. Guyatt GH, Oxman AD, Kunz R, et al. GRADE guidelines: 7. Rating the quality of evidence-inconsistency. *J Clin Epidemiol.* 2011;64(12):1294-1302.

52. Treadwell JR, Tregear SJ, Reston JT, Turkelson CM. A system for rating the stability and strength of medical evidence. *BMC Med Res Methodol.* 2006;6:52.

53. Guyatt GH, Oxman AD, Kunz R, et al. GRADE guidelines 6. Rating the quality of evidence-imprecision. *J Clin Epidemiol.* 2011;64(12):1283-1293.

54. Guyatt GH, Oxman AD, Kunz R, et al. GRADE guidelines: 8. Rating the quality of evidence-indirectness. *J Clin Epidemiol.* 2011;64(12):1303-1310.

55. Guyatt GH, Oxman AD, Montori V, et al. GRADE guidelines: 5. Rating the quality of evidence-publication bias. *J Clin Epidemiol.* 2011;64(12):1277-1282.

56. Guyatt GH, Oxman AD, Sultan S, et al. GRADE guidelines: 9. Rating up the quality of evidence. *J Clin Epidemiol.* 2011;64(12):1311-1316.

57. Tang KC, Choi BC, Beaglehole R. Grading of evidence of the effectiveness of health promotion interventions. *J Epidemiol Community Health.* 2008;62(9):832-834.

58. Baral SD, Wirtz A, Sifakis F, Johns B, Walker D, Beyrer C. The highest attainable standard of evidence (HASTE) for HIV/AIDS interventions: toward a public health approach to defining evidence. *Public Health Rep.* 2012;127(6):572-584.

59. Hill AB. The Environment and Disease: Association or Causation? *Proc R Soc Med.* 1965;58:295-300.

60. Watson SI, Lilford RJ. Integrating multiple sources of evidence: a Bayesian perspective. In: Raine R, Fitzpatrick R, Barratt H, Bevan G, Black N, Boaden R, et al. Challenges, solutions and future directions in the evaluation of service innovations in health care and public health. . *Health Services and Delivery Research.* 2016;4(16):1-18.

61. Guyatt G, Drummon R, Group E-BMW. *Users' guide to the medical literature: a manual for evidence-based clinical practice.* Chichago: American Medical Association; 2002.

62. Murad MH, Asi N, Alsawas M, Alahdab F. New evidence pyramid. *Evid Based Med.* 2016;21(4):125-127.

63. Geldsetzer P, Fawzi W. Experiments and Quasi-experiments: Complementary Approaches to Advancing Global Health Knowledge. *J Clin Epidemiol.* 2017.

64. Vandenbroucke JP, Broadbent A, Pearce N. Causality and causal inference in epidemiology: the need for a pluralistic approach. *Int J Epidemiol.* 2016;45(6):1776-1786.

65. Lawlor DA, Tilling K, Davey Smith G. Triangulation in aetiological epidemiology. *Int J Epidemiol.* 2016;45(6):1866-1886.

66. Hultcrantz M, Rind D, Akl EA, et al. The GRADE Working Group clarifies the construct of certainty of evidence. *J Clin Epidemiol.* 2017;87:4-13.

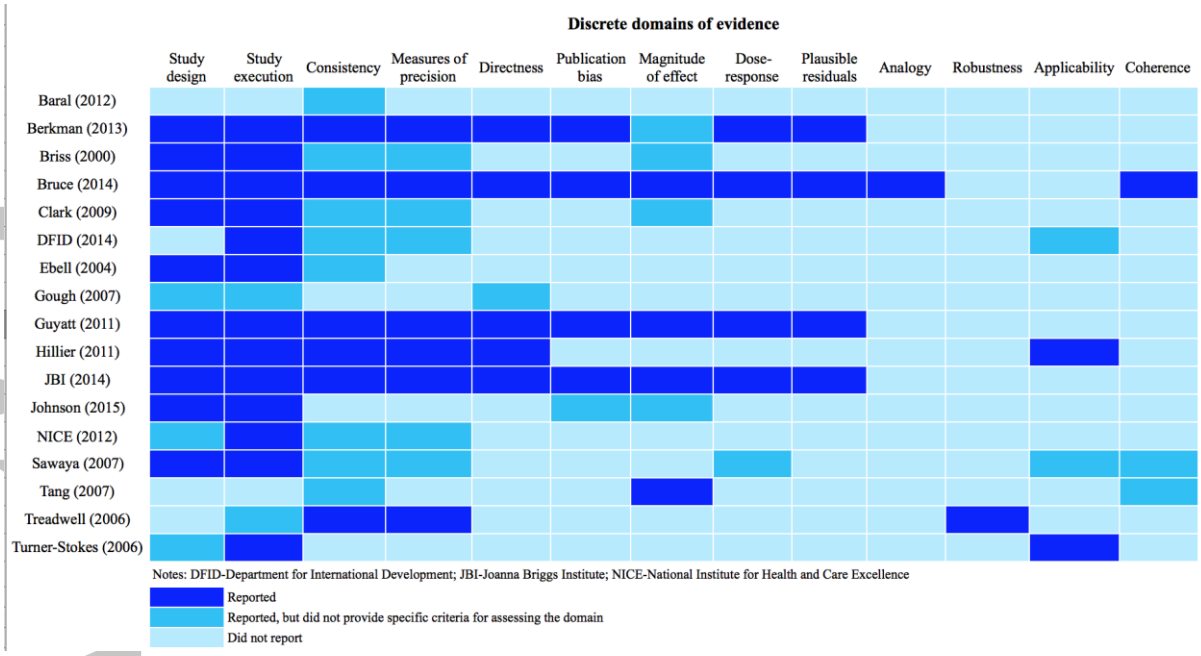

Figure 1. Reporting of the domains of evidence in the evidence rating systems for health and social interventions

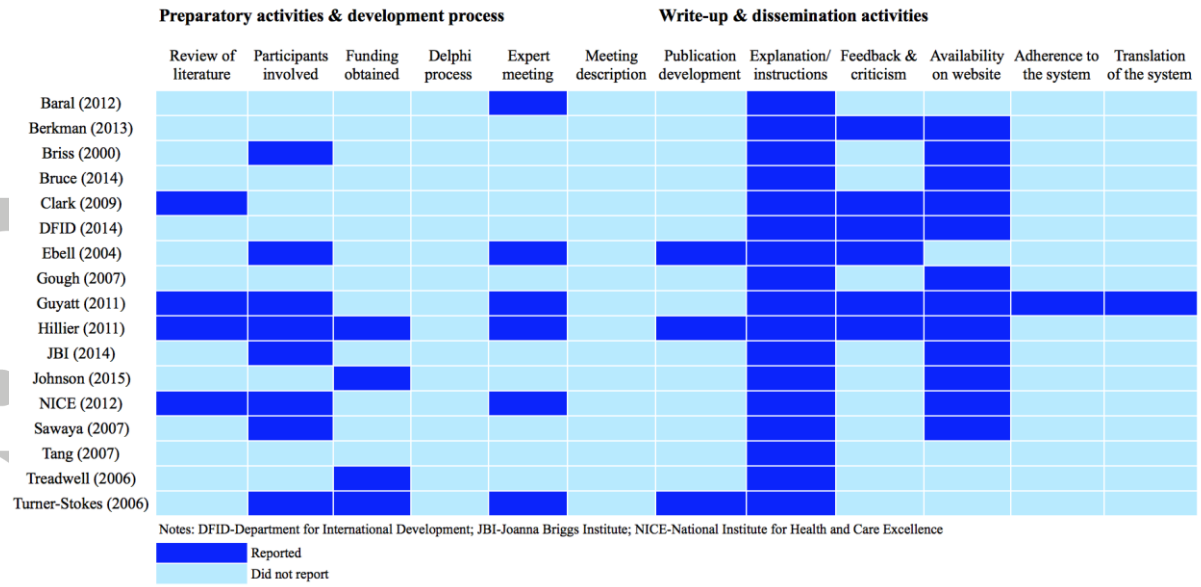

Figure 2. Reporting of the activities for developing and disseminating the evidence rating systems for health and social interventions

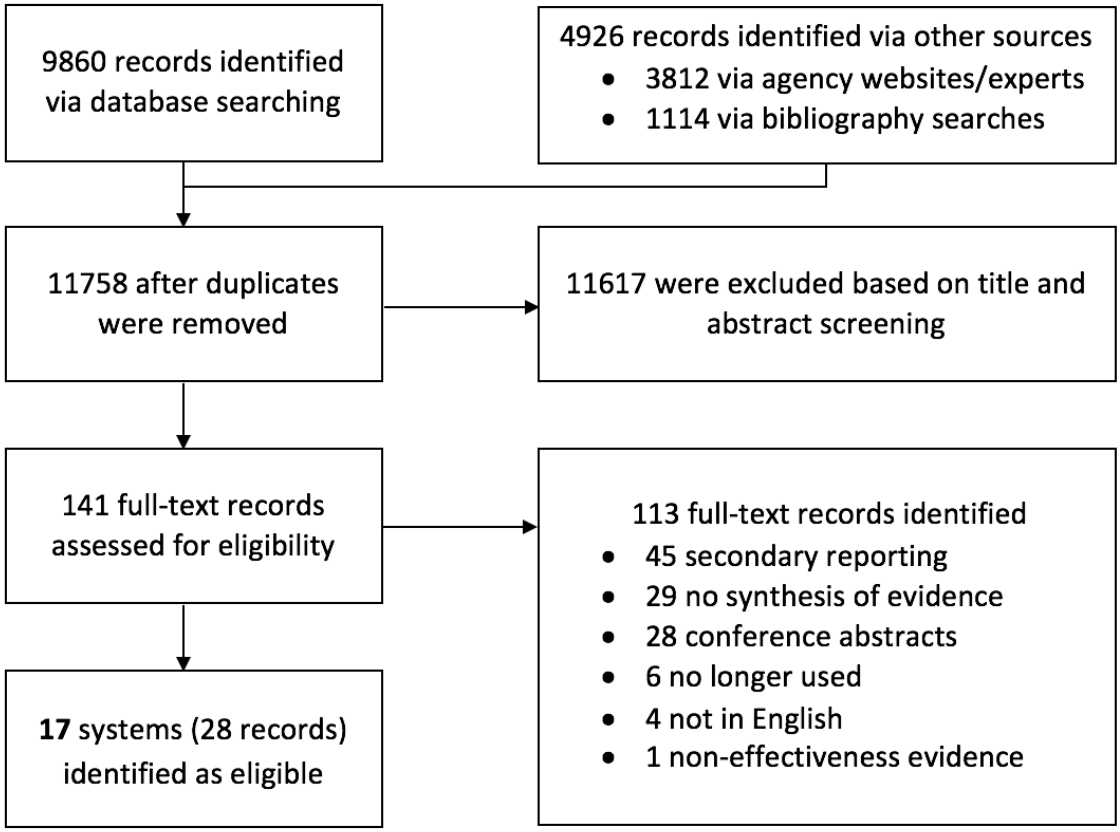

Figure 3. Systematic review PRISMA flow diagram

Table 1. Overview of the evidence rating systems for health and social interventions

First author (year) **Name of the system/organisation**	**Domains of evidence**	**Notes on the domains of evidence**	**Evidence ratings**	**Evidence synthesis approach**	**Context of application**
Baral (2012) ***The Highest Attainable Standard of Evidence (HASTE)*** *"…focuses on triangulation of three distinct categories of evidence" (p. 572)*	1. Efficacy data 2. Implementation data[a] 3. Plausibility data[a]	1. Consistent; inconsistent; limited 2. Consistent; inconsistent; limited[a] 3. High; low; undefined[a]	• Grade 1 (strong) • Grade 2 (conditional) • Grade 2a (probable) • Grade 2b (possible) • Grade 2c (pending) • Grade 3 (insufficient) • Grade 4 (inappropriate)	Not specified	Guideline development in public health (specific focus on HIV/AIDs interventions)
Berkman (2013) ***Agency for Healthcare Research and Quality (AHRQ)*** *"…confidence in systematic review conclusions so that decision-makers can use them effectively" (p. 1)*	1. Study Design 2. Study limitations 3. Directness 4. Consistency 5. Precision 6. Reporting bias 7. Dose-Response 8. Plausible confounding 9. Magnitude of effect 10. Applicability[a]	1. High (RCTs); Low (non-RCTs) 2. Risk of bias in RCTs/non-RCTs 3. Divergence from the outcomes & comparisons of interest 4. Consistency in magnitude or direction 5. Sample size, width of 95% CI 6. Publication bias; selective outcome reporting bias; selective analysis reporting 7. Dose-response relationship 8. Counteracting confounding 9. Size of the estimate of the effect 10. Likelihood of expected results under the "real-world" conditions[a]	• High • Moderate • Low • Insufficient	Quantitative: meta-analysis or narrative synthesis	Evidence synthesis in clinical medicine
Briss (2000) ***The Guide to Community Preventive Services*** *"…confidence that changes in outcomes are attributable to the interventions" (p. 38)*	1. Design suitability 2. Quality of study execution 3. Number of studies 4. Consistent 5. Effect size 6. Applicability[a] 7. Barriers to implementation[a] 8. Economic evaluations[a] 9. Other effects[a]	1. Greatest (concurrent comparison); moderate (comparison, but not concurrent); least (single group) 2. 6 categories of threats to validity: good, fair or limited 3. – 4. Consistent in direction & size 5. Defined on a case-by-case basis 6. Applicability to local situations[a] 7. – 8. – 9. Evidence on harms[a]	• Strong • Sufficient • Insufficient	Quantitative: meta-analysis or narrative synthesis	Guideline development in public health

Table 1. (Continued)

First author (year) Name of the system/organisation	Domains of evidence	Notes on the domains of evidence	Evidence ratings	Evidence synthesis approach	Context of application
Bruce (2014) ***Grading of evidence for public health interventions (GEPHI)*** *"…it is useful to make a distinction between: (a) strength of evidence for causal inference, for which Bradford Hill viewpoints for distinguishing causation from association in environmental epidemiology are often referred to…, and (b) the quality of evidence for the intervention effect size (confidence in the estimate), for which GRADE may be used" (p. 11)*	1. Study Design 2. Study limitations 3. Indirectness 4. Inconsistency 5. Imprecision 6. Reporting bias 7. Dose-Response 8. Plausible confounding 9. Magnitude of effect 10. Analogy 11. Consistency 12. Coherence	1. High (RCTs); Moderate (quasi-experimental designs); Low (other observational designs) 2. Risk of bias in RCTs/non-RCTs 3. Divergence from the PICO elements 4. Heterogeneity in the effect estimates 5. Sample size, width of 95% CI 6. Failure to identify studies 7. Dose-response relationship 8. Counteracting plausibility 9. Size of the estimate of the effect 10. Supporting evidence with similar mechanisms 11. Consistent evidence across different settings 12. Coherence in the overall causal chain: high, moderate, weak (separate rating)	• High • Moderate • Low • Insufficient	Quantitative: meta-analysis or narrative synthesis	Guideline development in public health
Clark (2009) ***Let Evidence Guide Every New Decision (LEGEND)*** *"The term 'level' was important to nurses to indicate the quality of an individual article; while 'grade' was more familiar to doctors and was adopted to indicate the quality of a body of evidence" (p. 1057)*	1. Study Quality 2. Consistency 3. Number of Studies	1. The aggregate quality ratings for individual studies (categorised based on an evidence hierarchy; e.g. 1a – good quality systematic review, 1b – lesser quality systematic review, 2a – good quality RCT/CCT; 2b – lesser quality RCT/CCT; etc.) 2. The extent to which similar findings are reported: Yes; No; Not available 3. –	• High • Moderate • Low • Grade-Not-Assignable	Not specified	Guideline development in clinical medicine
DFID (2014) ***Assessing the Strength of Evidence*** *"This Note assumes that the overall 'strength' of a body of evidence is determined by the "avoidance of bias" of studies that constitute it, and by the size, context and consistency" (p. 3)*	1. Quality 2. Size of the body of evidence 3. Consistency 4. Context of the body of evidence	1. Assessed with regards to 7 domains: high; moderate; low 2. Large; medium small 3. Consistent; inconsistent; mixed 4. Global; context-specific	• Very Strong • Strong • Medium • Limited • No evidence	Not specified	Evidence synthesis in international development

Table 1. (Continued)

First author (year) Name of the system/organisation	Domains of evidence	Notes on the domains of evidence	Evidence ratings	Evidence synthesis approach	Context of application
Ebell (2004) ***Strength of Recommendation Taxonomy (SORT)*** *"We use the term level of evidence to refer to individual studies. The strength (or grade) of a recommendation for clinical practice is based on a body of evidence" (p. 60)*	1. Study Quality 2. Consistency	1. Study Quality (combined with study design considerations based on an evidence hierarchy) 2. Consistent; inconsistent	• Level 1 (good quality) • Level 2 (limited quality) • Level 3 (other evidence)	Quantitative: meta-analysis	Guideline development in clinical medicine
Gough (2007) ***Weight of evidence: a framework for the appraisal of the quality and relevance of evidence*** *"Weight of evidence is a useful heuristic for considering how to make separate judgements on different generic and review specific criteria " (p. 11)*	1. Relevance of research design 2. Study execution 3. Relevance of the focus/context of evidence	1. A review specific judgement about the appropriateness of that form of evidence for answering the review question 2. Generally accepted criteria for evaluating the quality of evidence 3. A review specific judgement about the relevance of the focus of the evidence for the review question	• Weight of Evidence A • Weight of Evidence B • Weight of Evidence C • Weight of Evidence D	Not specified (different quantitative and qualitative approaches)	Evidence synthesis in education
Guyatt (2011) ***GRADE: grading of recommendations assessment, development and evaluation*** *"confidence that the estimates of the effect are correct" (p. 385)*	1. Study Design 2. Study limitations 3. Indirectness 4. Inconsistency 5. Imprecision 6. Publication bias 7. Dose-Response 8. Plausible confounding 9. Magnitude of effect	1. High (RCTs); Low (non-RCTs) 2. Risk of bias in RCTs/non-RCTs 3. Divergence from the PICO elements 4. Heterogeneity in effect estimates 5. Sample size, width of 95% CI 6. Failure to identify studies 7. Dose-response relationship 8. Counteracting confounding 9. Size of the estimate of the effect	• High • Moderate • Low • Very low	Quantitative: meta-analysis or narrative synthesis	Evidence synthesis & guideline development in clinical medicine & public health
Hillier (2011) ***FORM: An Australian method for formulating and grading recommendations in evidence-based guidelines*** *"…considering all of these elements across all of the research studies addressing the clinical question as a whole (the 'body of evidence')" (p. 2)*	1. Evidence Base 2. Consistency 3. Clinical Impact 4. Applicability 5. Generalisability	1. Quality; quantity and study design (evidence hierarchy: Level I – systematic reviews of RCTs; Level II – RCTs, Level III-1 – pseudorandomised trial; Level III-2 – comparative study with concurrent control; Level III-3 – comparative study without concurrent controls) 2. Excellent; good; poor 3. Very large; substantial; moderate; slight 4. Excellent; good; satisfactory; poor 5. Excellent; good; satisfactory; poor	• A (evidence trusted) • B (evidence mostly trusted) • C (some support) • D (weak evidence)	Not specified	Guideline development in clinical medicine and public health

Table 1. (Continued)					
Joanna Briggs Institute (2014) ***Levels of evidence and grades of recommendations*** *"One of the main reason for continuing with Levels of Evidence system is to assist in assigning GRADE pre-rankings…" (p. 4)*	1. Study Design 2. **The remaining domains of evidence follow those of the GRADE approach**	1. Level 1: experimental; Level 2: quasi-experimental; Level 3: Observational-Analytic; Level 4: Observational-Descriptive; Level 5: Expert opinion	• High • Moderate • Low • Insufficient	Quantitative: meta-analysis or narrative synthesis	Evidence synthesis in clinical medicine & public health
Johnson (2015) ***Introducing EMMIE: an evidence rating scale to encourage mixed-method crime prevention synthesis*** *"…in addition to considering the extent to which evaluations manage to rule out biases that might distort estimates of effect size, we also need to gauge the extent to which they contribute to understanding of the contexts/moderators" (p. 462)*	1. Effects 2. Mechanisms/Mediators[a] 3. Moderators/Contexts[a] 4. Implementation[b] 5. Economic analysis[a]	1. Consideration of evidence validity elements 2. Reference to and/or test of theory of change[a] 3. Reference to and/or analysis of data relating to pre-defined moderators[a] 4. Account of implementation or implementation challenges[a] 5. Estimation of marginal, total or opportunity costs[a]	Promotes descriptive profiles rather than a single overall score	Mixed-method synthesis	Evidence synthesis in criminology
National Institute for Health and Care Excellence (NICE, 2012) ***Methods for the development of NICE Public Health Guidance*** *"strength of evidence – reflecting the appropriateness of the study design to answer the question and the quality, quantity and consistency of evidence" (p. 89)*	1. Study design 2. Quality 3. Quantity 4. Consistency 5. Direction of the effect[a] 6. Size of the effect[a] 7. Applicability[a]	1. Appropriateness to answer the question 2. Assessment of both internal and external validity 3. – 4. – 5. Positive; Negative; Mixed; None[a] 6. Small; Medium; Large[a] 7. Applicability of evidence in terms of PICO elements[a]	• No evidence • Weak evidence • Moderate evidence • Strong Evidence • Inconsistent Evidence	Quantitative: meta-analysis or narrative synthesis Qualitative for questions other than intervention effectiveness	Guideline development in public health
Sawaya (2007) ***U.S. Preventive Services Task Force (USPSTF)*** *"The U.S. Preventive Services Task Force (USPSTF) defines certainty as "likelihood that the USPSTF assessment of the net benefit of a preventive service is correct" (p. 873)*	1. Study Design 2. Study Quality 3. Generalisability 4. Quantity 5. Consistency 6. Other	1. Evidence hierarchy (Level I – RCT; Level II-1 – controlled trial without randomisation; Level II-2 – cohort and case-control; Level II-2 – multiple time series; Level III - opinions) 2. Design specific: good; fair; poor 3. – 4. – 5. – 6. Dose-response; fit within a biologic model	Chain of evidence: • High • Moderate • Low	Not specified	Guideline development in clinical medicine
Table 1. (Continued)					

First author (year) Name of the system/organisation	Domains of evidence	Notes on the domains of evidence	Evidence ratings	Evidence synthesis approach	Context of application
Tang (2007) ***Grading of evidence of the effectiveness of health promotion interventions*** *"…the strength of evidence can be graded by using three criteria" (p. 832)*	1. Association 2. Repeatability 3. How it works	1. High (risk ratio of 2 or more) and statistically significant association: high, low, none 2. Reflects the consistency of the findings in different settings: wide, limited, none 3. Reflects the known cause-effect mechanism for the intervention under study: known; not known	• Grade 1 (strong) • Grade 2A (probable) • Grade 2B (possible) • Grade 2C (limited) • Grade 3 (insufficient)	Not specified	Evidence synthesis in public health
Treadwell (2006) ***A system for rating the stability and strength of medical evidence*** *"Our system draws a distinction between two types of conclusions: quantitative and qualitative…a quantitative conclusion characterises the size of the effect, whereas a qualitative conclusion characterises the direction of the effect" (p. 6)*	1. Quality 2. Quantity 3. Informativeness 4. Homogeneity 5. Robustness	1. High; moderate; low; very low 2. Criterion met; criterion not met (at least 3 studies and 80% having calculable effect sizes) 3. Effect Size 4. Homogeneous; heterogeneous 5. Tested through sensitivity analysis: robust; not robust	• Strong • Moderate • Weak • Inconclusive	Quantitative: meta-analysis	Evidence synthesis in clinical medicine
Turner-Stokes (2006) ***Generating the evidence base for the National Service Framework for long term conditions: a new research typology*** *"Each individual recommendation is then given an overall 'grade of research evidence' rating of A, B or C based on the quality of all the evidence supporting it and how much of it was directly relevant" (p. 97)*	1. Type of evidence 2. Study quality 3. Applicability	1. Primary research-based; secondary research-based; review-based (no classification based on an evidence hierarchy) 2. Quality is assessed on the basis of 5 questions to reach a max. score of 10 (includes a question on the appropriateness of the study design) 3. Population context of the study: direct; indirect	• GRADE A • GRADE B • GRADE C	Quantitative: meta-analysis or narrative synthesis Qualitative for questions other than intervention effectiveness	Evidence synthesis in clinical medicine & public health

[a]These domains of evidence go beyond rating the quality of a body of evidence on intervention effectiveness and are used in systems to further inform grading of the recommendations for practice. In the GRADE approach, these domains are separately specified as "Evidence to Decision" criteria (see Alonso-Coello et al., 2016).

Notes: CCT – controlled clinical trial; CI – confidence interval; DFID – Department for International Development; PICO – population, intervention, comparison, outcomes; RCT – randomised controlled trial;

Table 2. Approaches to defining certainty of evidence in GRADE (adapted from Hultcrantz et al., 2017)[66]

Setting	Degree of contextualisation	Threshold or range	How to set	What certainty rating represents
Primarily for systematic reviews and health technology assessment	Non-contextualised	Range: 95% CI	Using existing limits of the 95% CI	Certainty that the effect lies within the confidence interval
		OR≠1; RR≠1; HR≠1; RD≠0	Using the threshold of null effect	Certainty that the effect of one treatment differs from another
Primarily for systematic reviews and health technology assessment	Partially-contextualised	Specified magnitude of effect	E.g., small effect is the effect small enough to not use the intervention if adverse effects/costs are appreciable	Certainty in a specified magnitude of effect for one outcome (e.g., trivial, small, moderate or large)
Primarily for practice guidelines	Fully-contextualised	Threshold determined with consideration of all critical outcomes	Considering the range of effects on all critical outcomes, and the values & preferences	Confidence that the direction of the net effect will not differ from one end of the certainty range to the other

Notes: CI = "Confidence Interval". GRADE = "Grading of Recommendations Assessment, Development and Evaluation". HR = "Hazard Ratio". OR = "Odds Ratio". RD = "Risk Difference". RR = "Risk Ratio".