U.S. Advisory Committee on Immunization Practices (ACIP) Handbook for Developing Evidence-based Recommendations:

Formulating questions, conducting the systematic review, and assessing the certainty of the evidence using GRADE

Grading of Recommendations Assessment, Development and Evaluation (GRADE)

Last Updated April 22, 2024

Centers for Disease Control and Prevention (CDC)
Atlanta, GA, USA

Table of Contents

1. Introduction

The U.S. Advisory Committee on Immunization Practices (ACIP) provides expert external advice and guidance to the Director of the Centers for Disease Control and Prevention (CDC) and the Secretary of the Department of Health and Human Services (HHS) on the use of vaccines and related agents for control of vaccine-preventable disease in the U.S. civilian population.

Information on the charter, structure, role, processes, procedures, and membership of the ACIP are available at http://www.cdc.gov/vaccines/acip/committee/index.html.

The ACIP unanimously voted during the October 2010 meeting to adopt the Grading of Recommendations Assessment, Development and Evaluation (GRADE) approach for developing evidence-based recommendations[1]. The GRADE approach provides a framework for assessing the certainty (i.e., quality or confidence) of the evidence and moving from evidence to decision making (i.e., recommendations).

This handbook provides guidance to the ACIP workgroups on how to use the GRADE approach for assessing the certainty of evidence. The following sections include a brief overview of evidence retrieval and synthesis process to contextualize where the GRADE evidence assessment fits into the guideline development process.

Since 2010, there have been many advancements in the methods used to review the evidence and assess its certainty. This document replaces the 2013 version to include these updates and practical examples of all stages of the process.

Please refer to the ACIP Evidence to Recommendation User's Guide available at https://www.cdc.gov/vaccines/acip/recs/grade/about-grade.html for guidance on moving from evidence to decision-making, including whether GRADE should be used to address a policy question.

Resources and tools for implementing ACIP recommendations are available at http://www.cdc.gov/vaccines/.

References

1. Ahmed F, Temte JL, Campos-Outcalt D, Schünemann HJ, Group AEBRW. Methods for developing evidence-based recommendations by the Advisory Committee on Immunization Practices (ACIP) of the US Centers for Disease Control and Prevention (CDC). *Vaccine*. 2011;29(49):9171-9176.

2. Creating Trustworthy Guidelines

In 2011, the National Academy of Medicine, formerly Institute of Medicine, identified eight steps required to develop trustworthy guidelines[2]:

1. Establishing transparency
2. Managing conflict of interest
3. Composing the guideline development group
4. Providing for clinical practice guideline–systematic review intersection
5. Establishing evidence foundations for, and rating strength of, recommendations
6. Articulating recommendations
7. Obtaining external review
8. Updating regularly and with new evidence

This document outlines steps and strategies for conducting a rigorous systematic review and the methods for assessing the certainty of the evidence. The evidence, and the assessment of the certainty of evidence informs the strength of the recommendations, with the goal of maintaining transparency throughout the process.

References

2. Committee on Standards for Developing Trustworthy Clinical Practice Guidelines BoHCS, Institute of Medicine. *Clinical Practice Guidelines We Can Trust*. National Academies Press; 2011.

3. Overview of the Guideline Development Process

The typical guideline development process, including within ACIP, follows these steps:

1. Form the guideline development or work group panel*
2. Identify the scope of the guideline*
3. Formulate the research question(s) (i.e., Population, Intervention, Comparator, Outcomes [PICO]) and identify critical and important outcomes for decision-making to guide the process
4. Identify a high-quality published systematic review or conduct a systematic review *de novo*
5. Quantitatively synthesize and narratively summarize the results for all critical and important outcomes
6. Assess the body of evidence using GRADE and create GRADE evidence profiles
7. Complete the evidence-to-recommendation (EtR) decision-making framework
8. Formulate recommendations

*These steps may occur in tandem or in reverse

Figure 1. Diagram of the guideline development process[3]

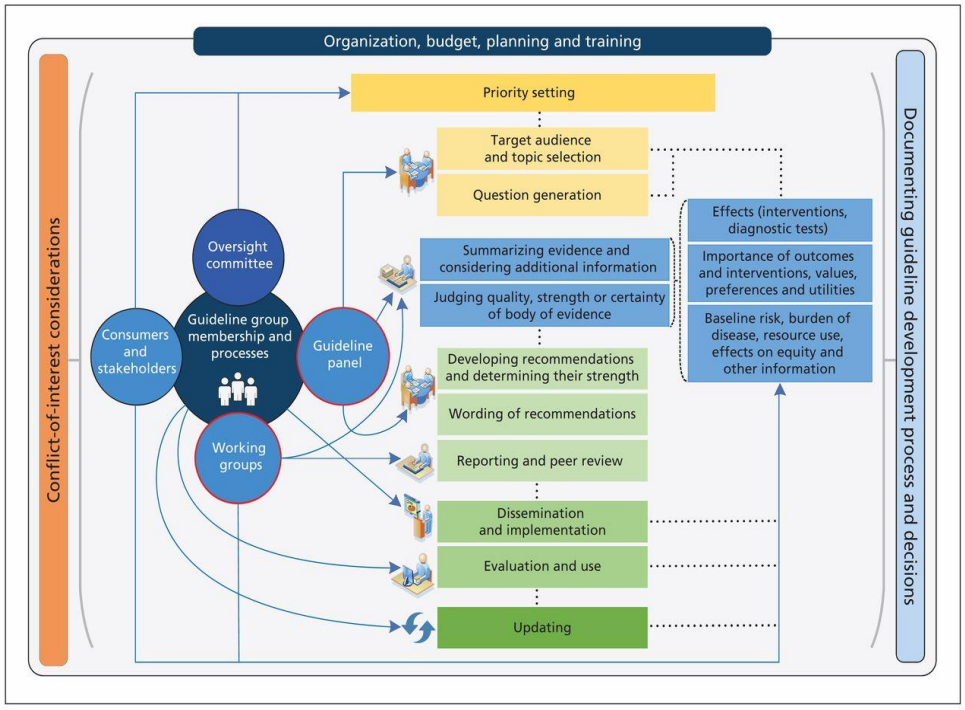


Figure 1 is not an ACIP-specific figure; the process may be individualized by organization.

References

3. Schünemann HJ, Wiercioch W, Etxeandia I, et al. Guidelines 2.0: systematic development of a comprehensive checklist for a successful guideline enterprise. *CMAJ*. 2014/02/18/ 2014;186(3):E123-E142. doi:10.1503/cmaj.131237

4. Formulating PICO Questions

Guidelines help answer questions about clinical, communication, organizational or policy interventions, in the hope of improving health care or health policy[4]. It is therefore helpful to structure a guideline in terms of answerable questions with relevant outcomes. Research questions that are too broad may necessitate extra resources to conduct the review, leading to heterogenous results that may be difficult to interpret. However, a broad question may produce a holistic summary based on a larger body of evidence and more generalizable findings. In contrast, narrow questions may require less resources but could lead to a smaller body of evidence with less generalizable findings. Depending on the scope of the review, authors should decide if it is more beneficial to lump things together resulting in a broader PICO question or if it is more useful to split comparisons and create narrow PICO questions[5]. To define the scope of the review, research priorities should be identified, and stakeholders should be engaged. The scope of the review may focus on a setting in which a vaccine is introduced where there was no vaccine previously or in a setting where a new vaccine is being compared to an existing vaccine. The scope may also be focused on new or updated recommendations for existing vaccines based on a changing epidemiology and/or the populations affected. For more information about developing a PICO question for a systematic review, refer to https://training.cochrane.org/handbook/current/chapter-02 [5].

The GRADE approach can be used to answer various questions that lead to actionable recommendations[6]. Research questions can be categorized as either background or foreground questions[4]. Background questions provide context and frame the need for the guideline, while foreground questions directly inform recommendations. Therefore, background questions can provide information on the prevalence or burden of a problem that help formulate foreground questions. Foreground questions provide insight on harms and benefits of the intervention of interest while also considering factors like acceptability and feasibility. Good questions target topics with controversy or doubt surrounding the answer, help pave the way for future research and positively impact patient care, costs and quality of life.

Research questions act as the starting point for formulating recommendations; they help inform inclusion and exclusion criteria for included studies, shape search strategies, inform the type of data to be extracted and guide the wording for recommendations[6]. Research questions may also inform the type of data synthesis used in a review. In order to create strong research questions that shape an evidence review, questions should be developed and presented using the PICO (population/intervention/comparator/outcome) framework[4]. The population component of the question describes the target population for the intervention. The intervention includes the treatment, test, policy or exposure being evaluated in the review. The comparator specifies the alternatives to the intervention being recommended in the guideline. Typically, a placebo is not used as a comparison in a recommendation (even though it may serve as a comparison in the systematic review of the literature), as “placebo” would not be a sensible option to recommend. Instead, the comparisons could look at existing alternatives, standard practice and no intervention (e.g., no vaccine in the situation when no

vaccine has been previously available). Outcomes consider the potential benefits and harms of the intervention and should be patient centered. Table 1 provides two examples of well-formulated PICO questions.

Table 1. Examples of PICO Questions

Policy question	**Example 1:** Should pre-exposure vaccination with the rVSVΔG-ZEBOV-GP vaccine be recommended for adults 18 years of age or older in the U.S. population who are at potential occupational risk of exposure to Ebola virus (species Zaire ebolavirus) for prevention of Ebola virus infection (*ACIP Grading for Ebola Vaccine | CDC, 2021*)?	**Example 2:** Should persons vaccinated with a MenB primary series who remain at increased risk for serogroup B meningococcal disease receive a MenB booster dose (*ACIP Grading for Serogroup B Meningococcal (MenB) Vaccines for Persons at Increased Risk for Serogroup B Meningococcal Disease | CDC, 2020*)?
Population	Adults aged 18 years or older in the United States who are at potential risk of exposure to EBOV because they are: ● Persons who are responding to an outbreak of EVD ● Persons who work as healthcare personnel (HCP) at federally designated Ebola Treatment Centers in the United States ● Persons who work as laboratorians or other staff at biosafety-level 4 facilities in the United States	Persons aged ≥10 years who have previously completed a MenB-FHbp or MenB-4C primary series who remain at increased risk for serogroup B meningococcal disease because of: • Persistent complement component deficiencies, complement inhibitor use, functional or anatomic asplenia, or routine exposure to isolates of Neisseria meningitidis as a microbiologist, or • An outbreak of serogroup B meningococcal disease
Intervention	Pre-exposure intramuscular immunization with a single licensed dose of the rVSVΔG-ZEBOV-GP vaccine	MenB-FHbp or MenB-4C booster dose
Comparison	No vaccine	No MenB-FHbp or MenB-4C booster dose
Outcomes	● Development of Ebola-related symptomatic illness ● Ebola-related mortality ● Vaccine-related joint pain or swelling (arthritis or arthralgia) ● Vaccine-related adverse pregnancy outcomes for women inadvertently vaccinated while pregnant and women who become pregnant within in 2 months of vaccination ● Transmissibility of rVSVΔG-ZEBOV-GP to humans or animals: Surrogate assessed with viral dissemination/shedding of the rVSVΔG-ZEBOV-GP vaccine virus ● Serious adverse events related to the vaccination ● Incidence and severity of oral or skin lesions ● Interaction or cross-reactivity with monoclonal antibody-based therapeutics or other VSV-backboned vaccines	● Serogroup B meningococcal disease ● Short-term immunogenicity of booster dose ● Persistence of immune response to booster dose ● Immune interference due to co-administration of booster dose with other vaccines ● Serious adverse events from booster dose

When developing PICO questions for guidelines, the setting in which the recommendations will be applied is often taken into consideration in addition to elements specified in the PICO approach. Moreover, the identification of subgroups within the population part of the PICO question allows for guideline panels to create recommendations targeting specific subpopulations. In the case of multiple comparators within a PICO question, guideline authors may need to clarify if the intervention is recommended over all the comparators equally or if there is a hierarchy. Additionally, PICO questions for guidelines may have a more comprehensive list of outcomes compared to a systematic review since they need to consider harms and how the implementation of an intervention may impact different populations. When listing outcomes for a PICO question informing a guideline document, the importance of the outcomes should be rated before the evidence review begins.

References

4. World Health O. *WHO handbook for guideline development*. World Health Organization; 2014:167.
5. Thomas J, Kneale D, McKenzie J, Brennan S, Bhaumik S. Chapter 2: Determining the scope of the review and the questions it will address. In: Higgins J, Thomas J, Chandler J, et al, eds. *Cochrane Handbook for Systematic Reviews of Interventions version 63 (updated February 2022)*. Cochrane; 2022. www.training.cochrane.org/handbook.
6. Guyatt GH, Oxman AD, Kunz R, et al. GRADE guidelines: 2. Framing the question and deciding on important outcomes. *J Clin Epidemiol*. 2011/04// 2011;64(4):395-400. doi:10.1016/j.jclinepi.2010.09.012

5. Choosing and Ranking Outcomes

In order to develop PICO questions for a systematic review, outcomes that are considered 'critical' or 'important' for decision-making to inform the recommendation need to be identified[4]. Outcomes should be selected based on relevance to the target population (i.e., who the guideline will benefit) and considerations of stakeholders. The outcome selection process should be conducted ideally during the protocol development stage and before the evidence is reviewed to avoid listing only outcomes measured in the existing literature; outcomes should focus on what might be critical and important in the decision-making process. Outcomes that are not addressed in the literature should not be disregarded because they may still influence the recommendation and help identify knowledge gaps. Searching the literature or surveying stakeholders can help to identify patient-important outcomes; however, it's not necessary, as these may also be informed by the guideline development panel or work group. The important aspect of identifying outcomes is that they are representative of those that the benefactors of the recommendation would be likely to weigh when deciding about whether or not to choose a specific course of action.

When evidence about a particular population-important outcome is very limited, surrogate outcomes may be considered to inform the health outcome. A surrogate or intermediate outcome refers to an indicator which serves as a measurement of a clinically meaningful outcome (e.g., anti-diphtheria toxoid antibody level as a surrogate outcome to the induction of immunity in individuals immunized against diphtheria). However, the population-important outcome for which the surrogate outcome is substituting for should be specified and considered when grading the certainty of the evidence. If a surrogate outcome is used, the importance of the corresponding health outcome (e.g., prevention of orthopoxviral disease) rather than that of the surrogate outcome (e.g., neutralizing antibody response or IgG levels) should be scored (Figure 2). The surrogate should not be listed as the health outcome of interest. Figure 2 shows a list of patient-important outcomes with surrogate outcomes mentioned separately to demonstrate that while they can be used when evidence is limited, they are not the outcome of interest. Use of a surrogate outcome requires assessment of the level of indirectness in informing the health outcome. When multiple surrogate outcomes are available, the least indirect to the health outcome should be assessed first to reduce the extent of potential indirectness[7].

After developing a list of relevant outcomes, the importance of these outcomes should be rated and ranked[8]. ACIP workgroup members should make an initial list of all possible relevant outcomes, including both desirable and undesirable effects. Each member should be asked to rate (score) the importance of each outcome on a 1 to 9 scale using a modified Delphi process, where 7–9 indicates that the outcome is critical for a decision, 4–6 indicates that it is important but not critical, and 1–3 indicates that it is of limited importance. Survey software (e.g., Google Forms, SurveyMonkey or Polls in Zoom) may be used to facilitate this process. The mean score for each outcome can be used to determine its relative importance, though it is helpful to provide the range of results as well, as this can give insight into any possible misunderstandings or divisions that warrant further discussion before finalizing the list

of outcomes. Figure 2 provides a visual representation of the scale[6]. Then the workgroup members will rank the highest rated outcomes so that the top 5–7 outcomes are included in the evidence review and evidence profile. It's worth noting that among the 5–7 outcomes included in the grading process, at least one must be an undesirable effect. Table 2a and 2b provides examples of outcome lists with initial rankings.

Figure 2. Theoretical example of Listing and Ranking Outcomes using the modified Delphi Process based on previously published guidelines for use of smallpox vaccine in laboratory and health-care personnel at risk for occupational exposure to orthopoxviruses [adapted][6,8,9]

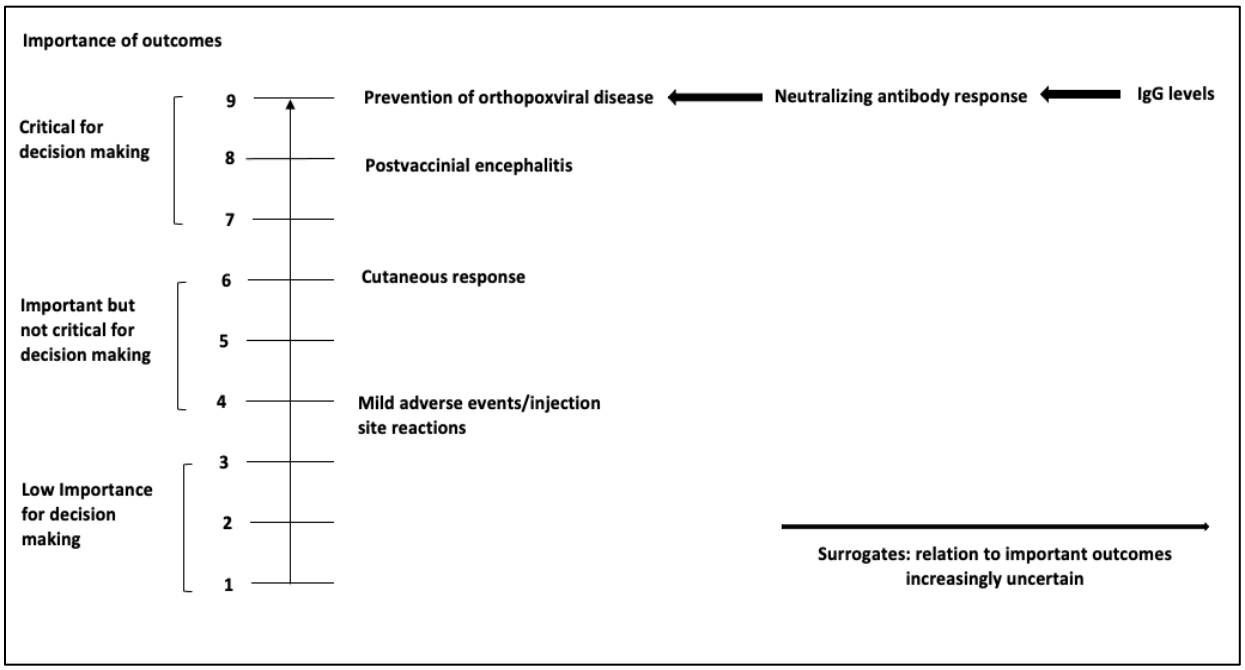


Table 2a. Example of Outcomes and Rankings[10]

Outcome	**Importance***	**Included in evidence profile (yes, no)**
Development of Ebola-related symptomatic illness	Critical for decision making	Yes
Ebola-related mortality	Critical for decision making	Yes
Vaccine-related joint pain or swelling (arthritis or arthralgia)	Critical for decision making	Yes
Vaccine-related adverse pregnancy outcomes for women inadvertently vaccinated while pregnant and women who become pregnant within in 2 months of vaccination	Critical for decision making	Yes
Transmissibility of rVSVΔG-ZEBOV-GP to humans or animals: Surrogate assessed with viral dissemination/shedding of the rVSVΔG-ZEBOV-GP vaccine virus	Critical for decision making	Yes
Serious adverse events related to the vaccination	Critical for decision making	Yes
Incidence and severity of oral or skin lesions	Not important for decision making	No
Interaction or cross-reactivity with monoclonal antibody-based therapeutics or other VSV-backboned vaccines	Not important for decision making	No

*Three options: 1. Critical for decision making; 2. Important but not critical for decision making; 3. Not important for decision making.

Table 2b. Examples of Outcomes and Ratings from "Use of JYNNEOS (orthopoxvirus) vaccine primary series for research, clinical laboratory, response team, and healthcare personnel (Policy Questions 1 and 2)"[11]

Outcome	**Importance***	**Included in Evidence Profile**
Prevention of disease (informed by geometric mean titer)**	Critical	Yes
Severity of disease	Important	Yes
Serious adverse events***	Critical	Yes
Myo-/peri-carditis	Critical	Yes
Minor adverse events	Not important	No

*Three options: 1. Critical for decision making; 2. Important but not critical for decision making; 3. Not important for decision making.
**Prevention of disease was informed by the surrogate outcome of geometric mean titer.
***Serious adverse events were defined according to the standard FDA definition. In addition, data was collected about any smallpox vaccine-specific adverse event: postvaccinal encephalitis, eczema vaccinatum, progressive vaccinia, and generalized vaccinia.

There are three points in time that the outcome rating can be updated as either critical, important or not important for decision-making. First, outcomes are listed and ranked before the systematic review search is conducted during the brainstorming phase. Second, the rating of the outcomes can be updated after the evidence has been retrieved if the literature provides rationale for why an outcome ranking needs to be altered. Lastly, when making the recommendation, the outcome ratings can be changed again if needed.

After the evidence review is conducted, outcome ranking can be reassessed because in some cases the importance of an outcome is better known after considering the literature. For instance:

- An outcome pertaining to a benefit may have been judged initially to be critical for making a recommendation, but it may no longer be considered to be critical if other benefits are evident, or;
- A suspected adverse event may be initially considered to be critical, but if the evidence review shows that the adverse event is not causally associated with the intervention, it may be considered important but not critical.

When creating evidence profiles, important and critical outcomes should be included even if no literature is available about them. The guideline recommendation will primarily be influenced by critical outcomes.

There may be a situation in which evidence, including surrogate or indirect evidence, is not identified to inform one or more of the critical or important outcomes that are set to be presented in the evidence

profile. One solution is to recognize there is no evidence to present and enter a "-" (dash) in each box within the evidence profile (including the certainty of evidence). In GRADEpro GDT, when editing the specific outcome there is an option "Not reported" which will autofill this for you. This transparently presents the results from the search and allows readers to recognize that future research may be needed to inform this outcome. If the work group feels that they can move forward with a recommendation in the absence of this specific outcome, the lack of evidence for this specific outcome will not influence the overall certainty of the evidence.

References

4. World Health O. *WHO handbook for guideline development*. World Health Organization; 2014:167.
6. Guyatt GH, Oxman AD, Kunz R, et al. GRADE guidelines: 2. Framing the question and deciding on important outcomes. *J Clin Epidemiol*. 2011/04// 2011;64(4):395-400. doi:10.1016/j.jclinepi.2010.09.012
7. Guyatt GH, Oxman AD, Kunz R, et al. GRADE guidelines: 8. Rating the quality of evidence--indirectness. *J Clin Epidemiol*. 2011/12// 2011;64(12):1303-1310. doi:10.1016/j.jclinepi.2011.04.014
8. Fitch K, Bernstein SJ, Aguilar MD, et al. *The RAND/UCLA Appropriateness Method User's Manual*. 2001. 2001/01/01/. Accessed 2022/03/06/21:27:33. https://www.rand.org/pubs/monograph_reports/MR1269.html
9. (ACIP) ACoIP. GRADE: Use of Smallpox Vaccine in Laboratory and Health-Care Personnel at Risk for Occupational Exposure to Orthopoxviruses. Centers for Disease Control and Prevention. https://www.cdc.gov/vaccines/acip/recs/grade/orthopoxvirus.html#t1
10. ACIP Grading for Ebola Vaccine | CDC. 2021/01/07/T05:56:55Z 2021;
11. (ACIP) ACoIP. Grading of Recommendations, Assessment, Development, and Evaluation (GRADE): Use of JYNNEOS (orthopoxvirus) vaccine primary series for research, clinical laboratory, response team, and healthcare personnel (Policy Questions 1 and 2). Centers for Disease Control and Prevention. 2024.

6. Systematic Review Overview

The evidence base must be identified and retrieved systematically before the GRADE approach is used to assess the certainty of the evidence and provide support for guideline judgements. A systematic review should be used to retrieve the best available evidence related to the PICO question. All guidelines should be preceded by a systematic review to ensure that recommendations and judgements are supported by an extensive body of evidence that addresses the research question. This section provides an overview of the systematic review process, external to the GRADE assessment of the certainty of evidence.

Systematic methods should be used to identify and synthesize the evidence[12]. In contrast to narrative reviews, systematic methods address a specific question and apply a rigorous scientific approach to the selection, appraisal and synthesis of relevant studies. A systematic approach requires documentation of the search strategy used to identify all relevant published and unpublished studies and the eligibility criteria for the selection of studies. Systematic methods reduce the risk of selective citation and improve the reliability and accuracy of decisions. The Cochrane handbook provides guidance on searching for studies, including gray literature and unpublished studies (Chapter 4: Searching for and selecting studies)[12].

6.1 Identifying the evidence

Guidelines should be based on a systematic review of the evidence[2,4]. A published systematic review can be used to inform the guideline, or a new one can be conducted. The benefits of identifying a previously conducted systematic review include reduced time and resources of conducting a review from scratch[4]. Additionally, if a Cochrane or other well-done systematic review exists on the topic of interest, the evidence is likely presented in a well-structured format and meets certain quality standards, thus providing a good evidence foundation for guidelines. As a result, systematic reviews do not need to be developed *de novo* if a high-quality review of the topic exists. Updating a relevant and recent high-quality review is usually less expensive and requires less time than conducting a review *de novo*. Databases, such as the Cochrane library, Medline (through PubMed or OVID), and EMBASE can be searched to identify existing systematic reviews which address the PICO question of interest. Additionally, the International Prospective Register of Systematic Reviews (PROSPERO) database can be searched to check for completed or on-going systematic reviews addressing the research question of interest[4]. It's important to base an evidence assessment and recommendations on a well-done systematic review to avoid any potential for bias to be introduced into the review, such as the inability to replicate methods or exclusion of relevant studies. Assessing the quality of a published systematic review can be done using the A Measurement Tool to Assess systematic Reviews (AMSTAR 2) instrument[4]. This instrument assesses the presence of the following characteristics in the review: relevancy to the PICO question, deviations from the protocol, study selection criteria, search strategy, data extraction process, risk of bias assessments for included studies and appropriateness of both

quantitative and qualitative synthesis[13]. A Risk of Bias of Systematic Reviews (ROBIS) assessment may also be performed[14].

If a well-done systematic review is identified but the date of the last search is more than 6-12 months old, consider updating the search from the last date to ensure that all available evidence is captured to inform the guideline. In a well-done published systematic review, the search strategy will be provided, possibly as an online appendix or supplementary materials. Refer to the Evidence Retrieval section (6.3) for more information.

If a well-done published systematic review is not identified, then a *de novo* systematic review must be conducted. Once the PICO question(s) have been identified, conducting a systematic review includes the following steps:

- Protocol development
- Evidence retrieval and identification
- Risk of bias assessment
- A meta-analysis or narrative synthesis
- Assessment of the certainty of evidence using GRADE

6.2 Protocol development

There are several in-depth resources available to support authors when developing a systematic review; therefore, this and following sections will refer to higher-level points and provide information on those resources. The Cochrane Handbook serves as a fundamental reference for the development of systematic reviews and the PRISMA guidance provides detailed information on reporting requirements. To improve transparency and reduce the potential for bias to be introduced into the systematic review process, a protocol should be developed *a priori* to outline the methods of the planned systematic review. If the methods in the final systematic review deviate from the protocol (as is not uncommon), this must be noted in the final review with a rationale. Protocol development aims to reduce potential bias and ensure transparency in the decisions and judgements made by the review team. Protocols should document the predetermined PICO and study inclusion/exclusion criteria without the influence of the outcomes available in published primary studies[15]. The Preferred Reporting Items for Systematic review and Meta-Analysis Protocols (PRISMA-P) framework can be used to guide the development of a systematic review [16]. Details on the PRISMA-P statement and checklist are available at http://www.prisma-statement.org/Extensions/Protocols[16]. If the intention is to publish the systematic review in a peer-reviewed journal separately from the guideline, consider registering the systematic review using PROSPERO before beginning the systematic review process[17].

To ensure the review is done well and meets the needs of the guideline authors, it is important to consider what type of evidence will be searched and included at the protocol stage before the evidence is retrieved[18]. While randomized controlled trials (RCTs) are often considered gold standards for evidence, there are many reasons why authors will choose to include nonrandomized studies (NRS) in their searches:

- To address baseline risks
- When RCTs aren't feasible, ethical or readily available
- When it is predicted that RCTs will have very serious concerns with indirectness (Refer to Table 12 for more information about Indirectness)

NRS can serve as complementary, sequential or replacement evidence to RCTs depending on the situation[19]. Section 9 of this handbook provides detailed information about how to integrate NRS evidence. At the protocol stage it is important to consider whether or not NRS should be included.

The systematic review team will scope the available literature to develop a sense of whether or not the systematic review should be limited to RCTs alone or if a reliance on NRS may also be necessary. Once this inclusion and exclusion criteria has been established, the literature can be searched and retrieved systematically.

6.3 Evidence retrieval and identification

6.3a. Searching databases

An expert librarian or information specialist should be consulted to create a search strategy that is applied to all relevant databases to gather primary literature[12]. The following databases are widely used when conducting a systematic review: MEDLINE (via PubMed or OVID); EMBASE; Cochrane Central Register of Controlled Trials (CENTRAL). The details of each strategy as actually performed, with search terms (keywords and/or MESH terms), the date(s) on which the search was conducted and/or updated, and the publication dates of the literature covered should be recorded.

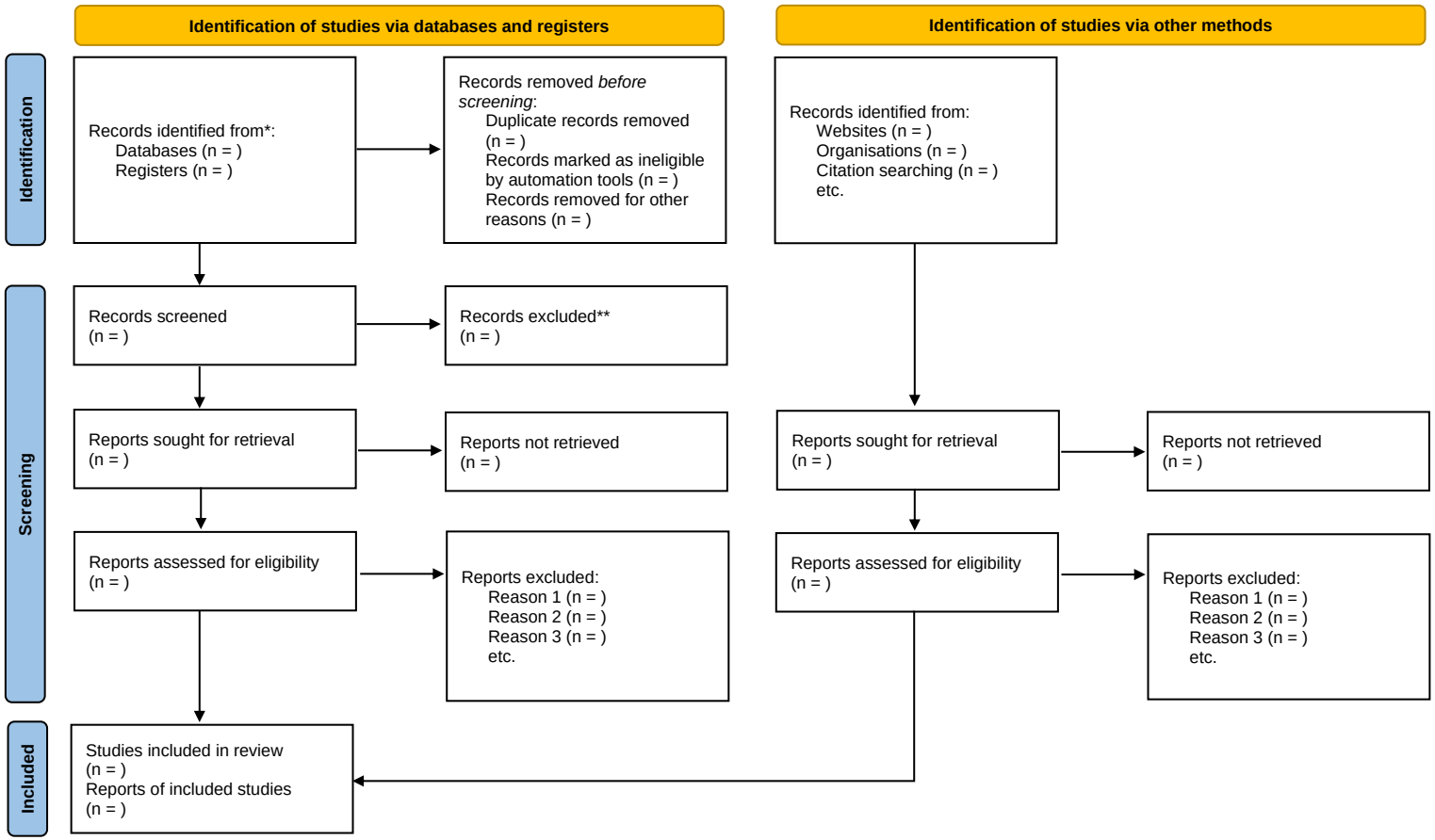
In addition to searching for evidence, references from studies included for the review should also be examined to add anything relevant missed by the searches. It is also useful to examine clinical trials registries maintained by the federal government (www.clinicaltrials.gov) and vaccine manufacturers, and consult subject matter experts. Ongoing studies should be recorded as well so that if the review or guideline were to be updated, these studies can be assessed for inclusion.

6.3b. Screening to identify eligible studies

The criteria for including/excluding evidence identified by the search, and the reasons for including and excluding evidence should be described (e.g., population characteristics, intervention, comparison, outcomes, study design, setting, language). Screening is typically conducted independently and in duplicate by at least two reviewers. Title and abstract screening is done first based on broader eligibility criteria and once relevant abstracts are selected, the full texts of those papers are pulled. The full-text screening is also usually conducted by two reviewers, independently and in duplicate with a more specific eligibility criteria to decide if the paper answers the PICO question or not. At both the title and abstract, and at the full-text stages, disagreements between reviewers can be resolved through discussion or involvement of a third reviewer. The goal of the screening process is to sort through the

literature and select the most relevant studies for the review. To organize and conduct the systematic review, Covidence can be used to better manage each of the steps of the screening process.. Other programs, such as DistillerSR or Rayyan can also be used to manage the screening process[20,21]. The PRISMA Statement (www.prisma-statement.org) includes guidance on reporting the methods for evidence retrieval. A PRISMA flow diagram (Figure 3) presents the systematic review search process and results.

Figure 3: PRISMA flow diagram depicting the flow of information through the different phases of the systematic review evidence retrieval process, including the number of records identified, records included and excluded at each stage, and the reasons for exclusions[22]



*Consider, if feasible to do so, reporting the number of records identified from each database or register searched (rather than the total number across all databases/registers).
**If automation tools were used, indicate how many records were excluded by a human and how many were excluded by automation tools.

From: Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021;372:n71. doi: 10.1136/bmj.n71. For more information, visit: http://www.prisma-statement.org/

6.3c. Data extraction

Once included articles have been screened and selected, relevant information from the articles should be extracted systematically using a standardized and pilot- tested data extraction form. Table 3 provides an example of an ACIP data extraction form (data fields may differ by topic and scope); Microsoft Excel can be used to keep track of and extract relevant details about each study. Data extraction forms typically capture information about: 1) study details (author, publication year, title, funding source, etc.); 2) study characteristics (study design, geographical location, population, etc.); 3) study population (demographics, disease severity, etc.); 4) intervention and comparisons (e.g., type of vaccine/placebo/control, dose, number in series, etc.); 5) outcome measures. For example, for dichotomously reported outcomes, the number of people with the outcome per study arm and the total number of people in each study arm are noted. In contrast, for continuous outcomes, the total number of people in each study arm, the mean or median, as well as standard deviation or standard error are extracted. This is the information needed to conduct a quantitative synthesis. If this information is not provided in the study, reviewers may want to reach out to the authors for more information or contact a statistician about alternative approaches to quantifying data. After extracting the studies, risk of bias should be assessed using an appropriate tool described in Section 8.1 of this handbook.

Table 3. Example of a data extraction form for included studies

Author, Year	Name of reviewer	Date completed	Study characteristics				Participants							Interventions		Outcomes	Other fields			
			Study design	Number of participants enrolled*	Number of participants analyzed*	Loss to follow up (for each outcome)	Country	Age	Sex (% female)	Race/ Ethnicity	Inclusion criteria	Exclusion criteria	Equivalence of baseline characteristics	Intervention arm Dose Duration Co-interventions	Comparison arm Dose Duration Co-interventions	Dichotomous: intervention arm n event/N, control arm n event/N Continuous: Intervention arm: Mean, SD, N, Control arm: Mean, SD, N	Type of study (published/ unpublished)	Funding source	Study period	Reported subgroup analyses

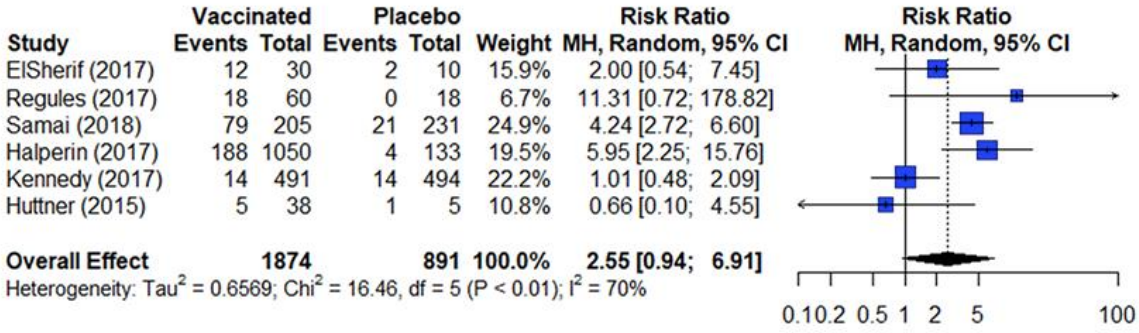
*total and per group

6.4 Conducting the meta-analysis

After the data has been retrieved, if appropriate, it can be statistically combined to produce a pooled estimate of the relative (e.g., risk ratio, odds ratio, hazard ratio) or absolute (e.g., mean difference, standard mean difference) effect for the body of evidence of each outcome. A meta-analysis can be performed when there are at least two studies that report on the same outcome. Several software programs are available that can be used to perform a meta-analysis, including R, STATA, and Review Manager (RevMan).

The results from a meta-analysis are presented in a forest plot as presented in figure 4. A forest plot presents the effect estimates and confidence intervals for each individual study and a pooled estimate of all the studies included in the meta-analysis[23]. The square represents the effect estimate and the horizontal line crossing the square is indicative of the confidence interval (CI; typically 95% CI). The area the square covers reflects the weight given to the study in the analysis. The summary result is presented as a diamond at the bottom.

Figure 4. Estimates of effect for RCTs included in analysis for outcome of incidence of arthralgia (0-42 days)[24]



The two most popular statistical methods for conducting meta-analyses are the fixed-effects model and the random-effects model[23]. These two models typically generate similar effect estimates when used in meta-analyses. However, these models are not interchangeable, and each model makes a different assumption about the data being analyzed.

A fixed-effects model assumes that there is one true effect size that can be identified across all included studies; therefore, all observed differences between studies are attributed to sampling error. The fixed-effect model is used when all the studies are assumed to share a common effect size[25]. Before using the fixed-effect model in a meta-analysis, consideration should be made as to whether the results will be applied to only the included studies. Since the fixed-effect model provides the pooled effect estimate for the population in the studies included in the analysis, it should not be used if the goal is to generalize the estimate to other populations.

In contrast, a random-effects model, some variability between the true effect sizes studies is accepted. These effect sizes are assumed to follow a normal distribution. The confidence intervals generated by the random-effects model are typically wider than those generated by the fixed-effect model, as they recognize that some variability in the findings can be due to differences between the primary studies. The weights of the studies are also more similar under the random-effects model. When variations in, for example, the participants or methods across different included studies is suspected, it is suggested to use a random-effects model. This is because the studies are weighed more evenly than the fixed-effect model. The majority of analyses will meet the criteria to use a random effects mode. One caveat about the selection of models: when the number of studies included in the analysis is few (<3), the random-effects model will produce an estimate of variance with poor precision. In this situation, a fixed-effect model will be a more appropriate way to conduct the meta-analysis[26].

References

2. Committee on Standards for Developing Trustworthy Clinical Practice Guidelines BoHCS, Institute of Medicine. Clinical Practice Guidelines We Can Trust. National Academies Press; 2011.
4. World Health Organization. *WHO handbook for guideline development, 2nd ed*. 2014: World Health Organization. 167.
12. Lefebvre C, Glanville J, Briscoe S, et al. Chapter 4: Searching for and selecting studies. In: Higgins J, Thomas J, Chandler J, et al, eds. Cochrane Handbook for Systematic Reviews of Interventions version 63 (updated February 2022). Cochrane; 2022. www.training.cochrane.org/handbook.
13. Shea BJ, Reeves BC, Wells G, et al. AMSTAR 2: a critical appraisal tool for systematic reviews that include randomised or non-randomised studies of healthcare interventions, or both. BMJ. 2017/09/21/ 2017:j4008. doi:10.1136/bmj.j4008
14. Bristol Uo. ROBIS tool.
15. Lasserson T, Thomas J, Higgins J. Chapter 1: Starting a review. In: Higgins J, Thomas J, Chandler J, et al, eds. Cochrane Handbook for Systematic Reviews of Interventions version 63. 2022. www.training.cochrane.org/handbook
16. Moher D, Shamseer L, Clarke M, et al. Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. Syst Rev. Jan 1 2015;4:1. doi:10.1186/2046-4053-4-1
17. PROSPERO. York.ac.uk. https://www.crd.york.ac.uk/PROSPERO/
18. Cuello-Garcia CA, Santesso N, Morgan RL, et al. GRADE guidance 24 optimizing the integration of randomized and non-randomized studies of interventions in evidence syntheses and health guidelines. J Clin Epidemiol. 2022/02// 2022;142:200-208. doi:10.1016/j.jclinepi.2021.11.026
19. Schünemann HJ, Tugwell P, Reeves BC, et al. Non-randomized studies as a source of complementary, sequential or replacement evidence for randomized controlled trials in systematic reviews on the effects of interventions. Research Synthesis Methods. 2013 2013;4(1):49-62. doi:10.1002/jrsm.1078
20. DistillerSR | Systematic Review and Literature Review Software. DistillerSR.
21. Rayyan – Intelligent Systematic Review. https://www.rayyan.ai/
22. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ. 2021;372:n71. doi:10.1136/bmj.n71
23. Deeks J, Higgins J, Altman D. Chapter 10: Analysing data and undertaking meta-analyses. In: Higgins J, Thomas J, Chandler J, et al, eds. Cochrane Handbook for Systematic Reviews of

Interventions version 63 (updated February 2022). Cochrane; 2022. www.training.cochrane.org/handbook.
24. Choi MJ, Cossaboom CM, Whitesell AN, et al. Use of ebola vaccine: recommendations of the Advisory Committee on Immunization Practices, United States, 2020. MMWR Recommendations and Reports. 2021;70(1):1.
25. Borenstein M, Hedges LV, Higgins JP, Rothstein HR. Introduction to meta-analysis. John Wiley & Sons; 2021.
26. Borenstein M, Hedges LV, Higgins JP, Rothstein HR. A basic introduction to fixed-effect and random-effects models for meta-analysis. Research Synthesis Methods. 2010;1:97-111. doi:DOI: 10.1002/jrsm.12

7. GRADE Criteria Determining Certainty of Evidence

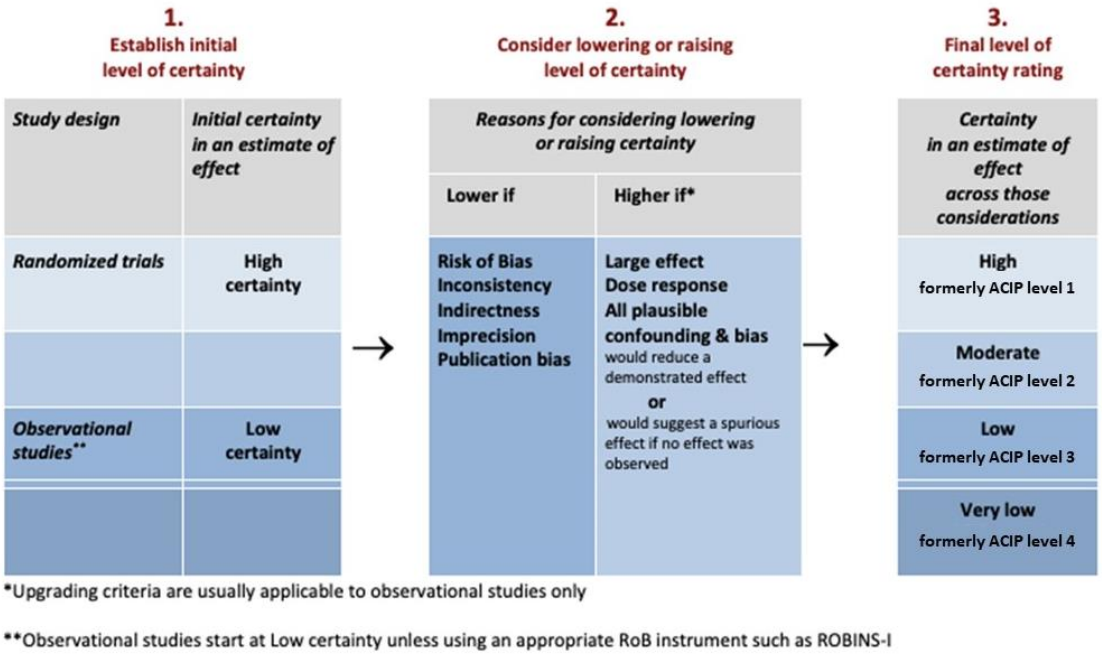
The GRADE approach is used to determine the certainty of evidence across the body of evidence for each outcome identified as critical or important for decision-making[6]. The certainty in the evidence reflects how confident we are that the observed effect reflects the true effect (Table 4).

The process of assessing the certainty of evidence begins by categorizing the study design into one of two groups:

- Randomized controlled trials (RCTs)
- Non-randomized studies (NRS) - also known as observational studies, i.e., cohort studies, case-control studies, controlled before-after studies, interrupted time series studies, and case series.

Randomized controlled trials initially start at a high level of certainty (former ACIP level 1) while non-randomized studies traditionally start at low level of certainty (former ACIP level 3) (Figure 5). This accounts for the lack of randomization in non-randomized studies, which increases the risk of residual or unknown confounding. However, if non-randomized studies are appropriately evaluated for risk of bias using a tool that assesses risk of bias along an absolute scale, such as the Risk Of Bias In Non-randomized Studies of Interventions (ROBINS-I) tool (currently available for comparative cohort studies), the evidence may start at an initial high certainty level[27]. The ROBINS-I tool assesses selection bias and confounding as an integral part of the evaluation process, unlike most other risk of bias tools for NRS[27]. The final certainty of evidence rating should not change based on the type of risk of bias instrument used. Five GRADE domains are used for downgrading the evidence type: risk of bias; inconsistency; indirectness; imprecision; and publication bias. Three GRADE criteria can be used to upgrade the evidence level of non-randomized studies: strength of association; dose-response; and opposing plausible residual confounding or bias. RCTs are typically not upgraded using these criteria as it risks erroneously inflating the certainty of the body of evidence.

Figure 5. GRADE criteria for assessing the type or certainty of evidence (adapted)[28]



The final "ACIP Level" certainty rating can be interpreted as how confident the authors are in the results. Formerly, these were ranked numerically (1—4) but ACIP has replaced numbers with the terms "high", "moderate", "low", "very low". Since older publications of GRADE will use the numerical levels, the correlates appear here for posterity. Table 5 presents the current and formerly used numerical ACIP levels of certainty in the evidence and how they can be conceptualized.

Table 4. Conceptualizing the certainty of the evidence[29]

High (formerly ACIP Level 1)	We are **very confident** that the true effect lies close to that of the estimate of the effect.
Moderate (formerly ACIP Level 2)	We are **moderately confident** in the estimate of effect: the true effect is likely to be close to the estimate of effect, but possibility to be substantially different.
Low (formerly ACIP Level 3)	Our **confidence** in the effect **is limited**: the true effect may be substantially different from the estimate of the effect.
Very low (formerly ACIP Level 4)	We have **very little confidence** in the effect estimate: the true effect is likely to be substantially different from the estimate of effect.

The final certainty of evidence for an outcome is cumulative of the considerations for rating down or rating up (non-randomized studies). For example, when the body of evidence from well-performed (i.e., no uncertainty or reason for rating down) NRS demonstrates both strength of association and dose-response, the evidence type may be rated up by two levels from Low to High (i.e., formerly ACIP Level 1). Typically, if the body of evidence for an outcome is rated down due to concerns from one or more of the previously described domains, it would not be rated up as this may overstate the certainty of an estimate thought to be substantially different from the truth. For example, if there is serious concern with the risk of bias due to lack of blinding, which may overestimate the effect, this outcome should not be rated back up due to large magnitude of effect.

Reviewers should categorize the final evidence certainty by making judgements on the individual GRADE domains in the context of their identified strengths or limitations. GRADE recognizes that judgment is involved during the evidence assessment and that overall certainty reflects if and how much concerns about the domains matter. It should be noted that concerns about domains for rating down may not equate in a one-to-one relationship to the overall certainty. For example, limitations pertaining to the risk of bias (e.g., the pooled analysis includes studies at both high and low risk of bias) and indirectness domains are identified, but these limitations are not serious enough for moving down each of the domains, the overall evidence type may be downgraded by one level when limitations for both domains are considered together (e.g., downgrade from high to moderate). The GRADE domain that played the biggest role in downgrading as well as all contributing factors should be specified.

The PICO question must be considered when determining the study design classification for an outcome. For example, a study in which infants are randomized into two different vaccination schedules would be classified as an RCT if the question is about which vaccination schedule is more effective. However, it would be classified as an NRS with no control group if the comparison group consists of infants who do not receive vaccination. Therefore, study design judgements should not be based on how authors of a study describe their methodology, but should consider how the study methodology aligns to answer the PICO question. This can be presented in the GRADE evidence profiles in one of two ways: 1) Identify study design as “Randomized Trial” to match the published study methodology and rate down twice for risk of bias with a footnote delineating that the evidence used to inform the outcome broke randomization; or 2) Identify study design as “Observational Study” and include a footnote that delineates the details of the trial. The PICO question should not be rephrased to reflect the evidence identified.

After conducting the GRADE assessment, the evidence can be categorized as either high, moderate, low or very low (formerly within ACIP, the equivalent levels were 1 [High], 2 [Moderate], 3 [Low], and 4 [Very low]). The certainty of the evidence reflects the confidence in the effect estimates that help inform recommendations. For guidelines, it is important to note that while the certainty of the evidence helps inform the recommendation, there are other factors that inform judgements about the strength of a recommendation. These can be found in the ACIP Evidence to Recommendation User’s Guide[30].

References

6. Guyatt, G.H., et al., *GRADE guidelines: 2. Framing the question and deciding on important outcomes.* J Clin Epidemiol. 2011 Apr;64(4):395-400. doi: 10.1016/j.jclinepi.2010.09.012. Epub 2010 Dec 30. PMID: 21194891.
27. Schünemann HJ, Cuello C, Akl EA, et al. GRADE guidelines: 18. 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. 2019/07// 2019;111:105-114. doi:10.1016/j.jclinepi.2018.01.012
28. Morgan RL, Thayer KA, Bero L, et al. GRADE: Assessing the quality of evidence in environmental and occupational health. Environ Int. 2016/08//Jul- undefined 2016;92-93:611-616. doi:10.1016/j.envint.2016.01.004
29. Schünemann HJ. Interpreting GRADE's levels of certainty or quality of the evidence: GRADE for statisticians, considering review information size or less emphasis on imprecision? J Clin Epidemiol. 2016/07// 2016;75:6-15. doi:10.1016/j.jclinepi.2016.03.018
30. ACIP Evidence to Recommendation User's Guide (Centers for Disease Control and Prevention) (2020).

8. Domains Decreasing Certainty in the Evidence

There are five GRADE domains for assessing limitations that can lower one's certainty in the evidence (i.e., the evidence level) for randomized trials and non-randomized studies (NRS): risk of bias (8.1); inconsistency (8.2); indirectness (8.3); imprecision (8.4); and publication bias (8.5).

8.1 Risk of bias (study limitations)

Study limitations may bias the estimates of the effect of an intervention on health outcomes[31]. The factors considered for evaluating study limitations or risk of bias (also referred to as internal validity) will depend on the study design. The number of studies is not a determining factor in determining risk of bias, as a single well-conducted study may result in high confidence in the estimated effect of vaccination on health outcomes. Risk of bias can differ amongst outcomes within an individual study, therefore, limitations for each outcome of interest in a study should be assessed separately.

Randomized Controlled Trials

For randomized controlled trials, Cochrane's revised risk of bias (RoB 2) tool can be used to assess study limitations[32,33]. The tool considers bias that may arise from the randomization process, deviations from the intended interventions, missing outcome data, measurement of the outcome and the selection of the reported result. Signaling questions are used to highlight concerns in each RoB domain. Judgements can express "High", "Low" or "Some concerns" with risk of bias. Details on how to use the tool and the various assessment questions can be found on the Risk of bias website[32]. Studies in which participants are allocated to intervention or control groups through quasi-randomization techniques (e.g., allocation by odd or even date of birth, date or day of admission, case record number, alternation/rotation) will automatically be at risk of selection bias due to inadequate generation of a randomized sequence, in addition to the ability of participants, or investigators enrolling participants, to foresee allocation[34]. Blinding outcome assessors is less important for the assessment of objective outcomes such as all-cause mortality, but is crucial for subjective outcomes such as quality of life. Risk of bias can differ across outcomes (e.g., higher risk of bias for subjective outcomes compared to objective outcomes when outcome assessors are not blinded; different subsets of studies for safety vs. efficacy studies). For adverse events or non-inferiority studies, intention-to-treat analyses may not be appropriate. If any information for assessing risk of bias is not reported in a publication, study investigators may be contacted. It may be possible to assess risk of bias from other reported information. For example, if information on allocation sequence concealment is not reported, data showing that the intervention and control groups are balanced at baseline may assuage concern regarding risk of bias. When assessing the risk of bias due to missing outcome data, reasons for the missing data and the quantity of missing data should both be taken into consideration. Table 5 provides a summary of the domains used in the RoB 2 assessment.

Table 5. Domains of RoB 2 tool

Study	Risk of bias arising from the randomization process (High/Low/Some Concerns)	Risk of bias due to deviations from the intended interventions (High/Low/Some Concerns)	Risk of bias due to missing outcome data (High/Low/Some Concerns)	Risk of bias in measurement of the outcome (High/Low/Some Concerns)	Risk of bias in selection of the reported result (High/Low/Some Concerns)

The Cochrane group has also developed risk of bias assessment tools to use for cluster-randomized trials and crossover trials[32] .

Non-randomized Studies

The criteria for assessing non-randomized studies like cohort studies, case-control studies, controlled before-after studies, interrupted time series, and case series differs from risk of bias assessments for randomized trials[31]. The Cochrane group recommends using the Risk Of Bias In Non-randomized Studies of Interventions (ROBINS-I) tool to assess the risk of bias for non-randomized studies, specifically for comparative cohort studies[35]. Similar to the RoB 2 tool recommended for RCTs, ROBINS-I assessments are done for specific results; each reported outcome study should be considered separately rather than judging the study as a whole. Confounding and co-interventions are major concerns that could lead to bias in non-randomized studies. Other domains such as selection bias, information bias, and reporting bias are also evaluated using the ROBINS-I tool; details on the signaling questions and domains used in the tool can be found on the Risk of bias website.

Table 6 provides an overview of the domains used in the ROBINS-I tool. Each domain is judged to have "Low", Moderate", or "Critical" risk of bias. "No information (NI)" is used when there is insufficient information to make a judgment on a domain. When using this tool, NRS start off with high certainty and can be graded down for study limitations after the ROBINS-I tool is used and concerns with risk of bias are identified[27]. The ROBINS-I tool uses an absolute metric rather than comparing non-randomized studies to a standard ideal NRS, thus making it easier to compare RCTs and non-randomized studies, as both are assessed using a similar metric for risk of bias.

Table 6. Domains of the ROBINS-I tool for NRS

Study	Bias due to confounding (Low/Moderate/Critical/NI)	Bias in selection of participants into the study (Low/Moderate/ Critical/NI)	Bias in classifications of interventions (Low/Moderate/ Critical/NI)	Bias due to deviations from intended interventions (Low/Moderate/ Critical/NI)	Bias due to missing data (Low/Moderate/ Critical/NI)	Bias in measurement of outcomes (Low/Moderate/ Critical/NI)	Bias in selection of the reported result (Low/Moderate/ Critical/NI)

The Newcastle-Ottawa Scale (NOS) is another tool that has been developed to assess the risk of bias of nonrandomized studies[36].

After using a tool to assess the risk of bias for each outcome in an individual study, the extent of study limitations for the body of evidence is categorized into one of the following groups[31]:

- No serious limitations (do not downgrade evidence type): most of the studies comprising the body of evidence have low risk of bias for all key criteria for evaluating study limitations.
- Serious limitations (downgrade one level): most of the studies have crucial limitations for one criterion or some limitations for multiple criteria that lower confidence in the estimated effect of vaccination on the outcome of interest.
- Very serious limitations (downgrade two levels): most of the studies have crucial limitations for one or more criteria that substantially lower confidence in the estimated effect.
- Extremely serious limitations (downgrade three levels)[37]: most of the studies have crucial limitations for multiple criteria that substantially lower confidence in the estimated effect. This option exists only for studies which are evaluated using ROBINS-I tool. The use of ROBINS-I here starts the evidence at high certainty.

When considering a body of evidence in which some studies have no serious limitations, some have serious limitations, and some have very serious limitations, it is not appropriate to automatically assign an average rating of serious limitations for the group of studies. When the risk of bias varies across studies, principles for determining whether to downgrade the evidence type for a group of studies include[31]:

- Consider the extent to which each study contributes to the overall or pooled estimate of effect. Larger studies with many outcome events will contribute more.
- Assess whether the results differ for studies with low risk of bias and those with high risk of bias. Consider focusing on studies with lower risk of bias if the results differ by risk of bias.
- Downgrade when there is substantial risk of bias across most of the studies.
- Consider limitations pertaining to the other GRADE criteria (if there are close calls regarding risk of bias with another GRADE criterion, consider downgrading the evidence level for at least one of the two GRADE criteria)

When close-call situations occur, this should be made explicit, and the reason for the ultimate classification should be stated. Table 7a provides an example of when results from NRS may not have serious concerns with risk of bias, while the body of evidence consisting of randomized trials has concerns with study limitations. Since the trials used subjective reporting of the outcome and lacked blinding, the body of evidence was downgraded due to serious concerns with risk of bias.

Table 7b presents a situation in which the certainty of the evidence from RCTs for the outcomes of serious adverse events and myo-/pericarditis were judged as very low; therefore, the work group considered the evidence from NRS. For both of these outcomes, the RCTs had concerns due to the small number of events and total patients. The NRS provided complementary evidence with a larger number of participants and results consistent with those from RCTs.

Table 7a. Evidence profile for outcome of incidence of arthritis (5–56 days)[24]

Certainty assessment							No. of patients		Effect		Certainty	Importance
No. of studies	**Study Design**	**Risk of Bias**	**Inconsistency**	**Indirectness**	**Imprecision**	**Other considerations**	**rVSV-vaccine**	**No rVSV-vaccine**	**Relative (95% CI)**	**Absolute (95% CI)**		
4	Randomized trials	Serious[a]	Not serious	Not serious	Serious[b]	None	39/1776 (2.2%)	16/868 (1.8%)	RR 1.80[d] (0.21 to 15.13)	23 more per 1,000 (from 22 fewer to 400 more)	Low	Critical
2	Non-randomized studies	Not serious	Not serious	Not serious	Very serious[b,c]	None	43/520 (8.3%)	3/107 (2.8%)	RR 2.06[d] (0.0001 to 7739.16)	33 more per 1,000 (from 28 fewer to 1000 more)	Very Low	Critical

Note: Non-randomized studies without comparators are not included in evidence table, but would be considered to offer very low certainty (evidence type 4)

Explanations

a. Studies used variable definitions and methods for diagnosing and reporting arthritis. In addition, participants, healthcare personnel, and outcome assessors were not blinded in Huttner 2015 or Samai 2018 potentially influencing events reported for this subjective outcome.

b. The 95% CI includes the potential for possible harms, as well as possible benefit.

c. Few events reported do not meet optimal information size and suggest fragility in the estimate.

d. RR calculated using the standard continuity correction of 0.5 and the overall effect uses a random effects model.

Table 7b. Evidence profile for Use of JYNNEOS (orthopoxvirus) vaccine heterologous for those who received ACAM2000 primary series[38]

Certainty assessment							№ of patients		Effect		Certainty	Importance
№ of studies	Study design	Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations	a booster dose of JYNNEOS	a booster dose of ACAM2000	Relative (95% CI)	Absolute (95% CI)		
A. Prevention of disease (assessed with: seroconversion rate)												
3 [1,2,3,4,5,6,7]	observational studies	serious[a]	not serious	serious[b]	serious[c]	none	No comparison data available. Intervention data from the systematic review: 272/333 (81.68 %) participants from 3 studies seroconverted 14 days after booster with MVA.				VERY LOW	CRITICAL
B. Severity of disease (assessed with: take maximum lesion area)												
1 [8]	observational studies	serious[a,d]	not serious	not serious	very serious[e]	none	No comparison data available. Intervention data from the systematic review: 20/20 (100%) of vaccinia experienced participants developed an attenuated take lesion after Dryvax challenge following booster with MVA vaccine.				VERY LOW	IMPORTANT
C. Serious adverse events (assessed with: vaccine related serious adverse event rate)												
1 [8]	randomized trials	serious[f]	not serious	not serious	very serious[g]	none	0/22 (0.0%)	0/28 (0.0%)	not estimable		VERY LOW	CRITICAL
C. Serious adverse events (assessed with: vaccine related serious adverse event rate)												
4 [1,2,3,4,5,6,7,9]	observational studies[h]	not serious	not serious	serious[i]	very serious[g]	none	0/367 (0.0%)[j]	3/1371 (0.2%)[k]	RR 0.53 (0.03 to 10.32)	1 fewer per 1,000 (from 2 fewer to 22 more)	VERY LOW	CRITICAL
D. Myo-/pericarditis (assessed with: myo-/pericarditis event rate)												
1 [8]	randomized trials	very serious[l]	not serious	not serious	very serious[m]	none	0/22 (0.0%)	0/28 (0.0%)	not estimable		VERY LOW	IMPORTANT
D. Myo-/pericarditis (assessed with: myo-/pericarditis event rate)												

Certainty assessment							№ of patients		Effect		Certainty	Importance
№ of studies	Study design	Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations	a booster dose of JYNNEOS	a booster dose of ACAM2000	Relative (95% CI)	Absolute (95% CI)		
3 [1,2,3,4,5,6,7]	observational studies	not serious	not serious	serious[i]	very serious[m]	none	0/349 (0.0%)[n]	0/1371 (0.0%)[o]	not estimable		VERY LOW	IMPORTANT

RR: risk ratio; **CI:** confidence interval

Explanations

a. Risk of bias due to lack of comparison data.
b. Seroconversion rate is an indirect measure of prevention.
c. Small sample size, no comparison.
d. Attrition rate was variable across study groups. One group lost 17% of participants.
e. Small sample size, fragility of estimate.
f. In the protocol it is unclear how serious adverse events were assessed.
g. Sample size is small, too small to detect rare adverse events.
h. Observational data was included in the evidence profile for this outcome because the effect estimate for the randomized trials was not estimable.
i. Single-arm studies contribute data to the intervention, but no available data for the comparison from the systematic review. Downgraded for indirectness because historical data was used for comparison.
j. Intervention data was drawn from 3 observational studies included in the systematic review. 0/349 (0.00 %) participants from 3 studies developed vaccine related serious adverse events.
k. Comparison data was drawn from historical data. In a phase III clinical trial for ACAM2000 enrolling participants with previous smallpox vaccination 3/1371 (0.22%) developed vaccine related serious adverse events after ACAM2000 administration. No smallpox vaccine-specific serious adverse event was recorded.
l. Assessment of myo-/pericarditis was initiated late in the study at the request of FDA. Very few subjects could be evaluated at that point. It was unclear how many subjects were evaluated.
m. Sample size is small, too small to detect rare events of myopericarditis after JYNNEOS®.
n. Intervention data was drawn from 3 observational studies included in the systematic review. 0/349 (0.00 %) participants developed myo-/pericarditis.
o. Comparison data was drawn from historical data. In a phase III clinical trial for ACAM2000 enrolling participants with previous smallpox vaccination, 0/1371 (0.00%) developed myo-/pericarditis after ACAM2000 administration.

8.2 Inconsistency

Inconsistency refers to an unexplained heterogeneity in the effect estimates across studies contributing to a summary estimate (e.g., relative risk or odds ratio for binary outcomes; mean difference for continuous outcomes)[39]. Inconsistency can be assessed by examining the following indicators of heterogeneity: 1) visual examination of the forest plot (point estimates and confidence intervals); 2) calculating statistical test of heterogeneity])- Chi-squared (Chi^2 or X^2) statistic; 3) calculating the (I-squared [I^2]; 4) contextualizing the findings with the target for our certainty rating.

Heterogeneity occurs when there is large variability between the studies pooled in a meta-analysis. Visual inspection can show effects that differ from the rest and should include an examination of the point estimates and overlap of confidence intervals[40]. A forest plot suggesting heterogeneity would show confidence intervals from individual studies that have limited or no overlap with the summary estimate. The studies contributing to the summary estimate may have point estimates that widely differ. However, difference may not only be detected by visualization; therefore, complementing this with numerical estimates of heterogeneity may be helpful. The I^2 statistic describes the percentage of variation across studies that is due to heterogeneity rather than chance. The higher the I^2 statistic, the more likely the variability seen is due to more than just change (I^2 >30% is low, ~50% is moderate, and >75% is substantial and requires further exploration). The Chi^2 tests the null hypothesis that the included studies are not different (homogenous); however, the results are susceptible to studies with small samples or if there are few studies in the meta-analysis. If the Chi^2 is small and the p-value large (>0.10 or >0.05; i.e., not significant) heterogeneity may not be suspected. Lastly, if the point estimate of the pooled estimate visually falls within the 95% CI of the studies included in the analysis, heterogeneity is less of a concern.

When making decisions about the extent to which heterogeneity contributes to our certainty rating (i.e., should we rate down for inconsistency and by how much), the target (threshold or range) of our certainty rating must be identified[41]. This could be the null, a minimally important difference, a range of magnitudes of trivial, small, moderate or large. Inconsistency is a concern when it crosses possible thresholds of meaning. Inconsistency may not be a concern when all of the point estimates (and CIs) of included studies lie above a given threshold *even if* they are disparate (e.g., visually confidence intervals don't overlap or I^2 is high, etc.).

In addition to noting the presence of inconsistency, it is desirable to determine potential reasons for the inconsistency. Differences in the following may result in inconsistency:

- Populations (e.g., vaccines may have different relative effects in sicker populations);
- Interventions (e.g., different effects with different number of doses or comparators);
- Outcomes (e.g., duration of follow-up);
- Study methods (e.g., studies with higher and lower risk of bias).

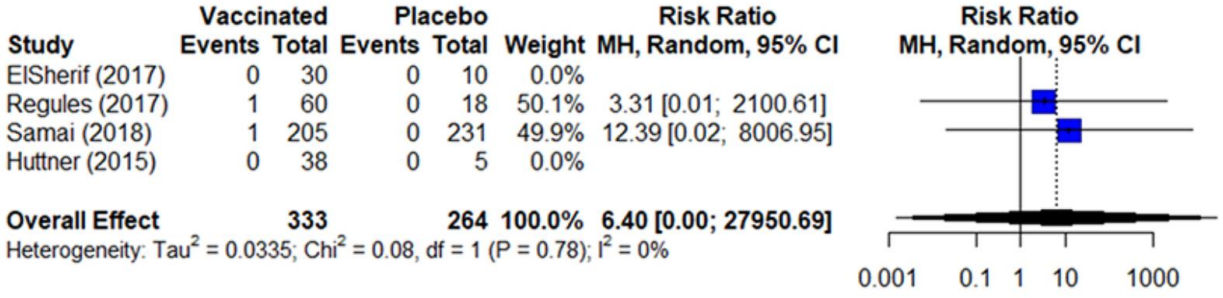
When heterogeneity is large and a plausible explanation cannot be identified, the evidence level should be downgraded by one or two levels, depending on heterogeneity in the magnitude of effect. While there are not specific guidelines for this; see "GRADE guidelines: 7. Rating the quality of evidence—inconsistency" for examples of downgrading[39]. If inconsistency can be explained, estimates of effect should be presented separately for the stratification that explains the observed heterogeneity. If results differ by study methods, preference may be given to results of studies with a lower risk of bias. If results differ by population groups, different recommendations may be made for different groups. If only one study is available, there are by default no concerns with inconsistency (i.e., select "Not serious" when grading).

Inconsistency is assessed more strictly in binary/dichotomous outcomes (relative values) than continuous outcomes (absolute values). For binary outcomes, inconsistency should be assessed using risk ratio or odds ratio which are measures of relative effect, where a value of 1 indicates the estimated effect is similar for both the intervention and comparison group[42]. Conversely, the risk difference is a measure of absolute effect that represents the difference in the observed risk and should not be used to assess inconsistency because it is very sensitive to the baseline risk (i.e., risk in control group) and baseline risk can differ substantially between studies[39]. The forest plot below (Figure 6) shows four studies included in the analysis for the binary outcome of severe (grade 3) arthralgia. Here, two studies contribute to the effect estimate (risk ratio), as they contain events. Visually, the pooled estimate (6.40) falls within the 95% CIs of the included studies; the Chi^2 is small (0.08) and the p-value is large (i.e., not significant at 0.10), and the $I^2 = 0\%$[24]. Based on all three steps, heterogeneity is not serious for this outcome.

To recap, any of the following factors may result in rating down for inconsistency:

1. I^2 is large (I^2 >30% is low, ~ 50% is moderate, and >75% is substantial and requires further exploration).
2. Statistical test for heterogeneity (Chi^2) shows a low P-value (i.e., < 0.05).
3. Confidence intervals of the point estimates of included studies do not overlap or show minimal overlap.

Figure 6. Estimates of effect for RCTs included in analysis for outcome of incidence of severe (grade 3) arthralgia (0-42 days) [24]



Effect estimates from continuous outcomes can be presented in a number of ways. If the primary studies included have assessed an outcome using the same scale, then it can be presented as a Mean Difference (MD). However, when pooling studies which measure the same continuous outcome using

different instruments or varying scales, researchers might choose to present this as a Standardized Mean Difference (SMD). The MD can be easily interpreted and assessed for heterogeneity and inconsistency. However, SMD might pose more of a difficulty and reviewers might need to use a different approach to further present and interpret the effect estimate[43]. Tables 8 and 9 present the options available to reviewers dealing with studies with these challenges.

Table 8: Five approaches to presenting results of continuous variables when primary studies have used different instruments to measure the same construct[43]

Approach	Advantages	Disadvantages	Recommendation
SD units (standardized mean difference; effect size)	Widely used	Interpretation challenging Can be misleading depending on whether population is very homogenous or heterogeneous	Do not use as the only approach
Present as natural units	May be viewed as closer to primary data	Few instruments sufficiently used in clinical practice to make units easily interpretable	Approaches to conversion to natural units include those based on SD units and rescaling approaches. We suggest the latter. In rare situations when instrument very familiar to frontline clinicians, seriously consider this presentation
Relative and absolute effects	Very familiar to clinical audiences and thus facilitate understanding Can apply GRADE guidance for large and very large effects	Involve assumptions that may be questionable (particularly methods based on SD units)	If the MID is known, use this strategy in preference to relying on SD units Always seriously consider this option
Ratio of means	May be easily interpretable to clinical audiences Involves fewer questionable assumptions than some other approaches Can apply GRADE guidance for large and very large effects	Cannot be applied when measure is change and therefore negative values possible Interpretation requires knowledge and interpretation of control group mean	Consider as complementing other approaches, particularly the presentation of relative and absolute effects
MID units	May be easily interpretable to audiences Not vulnerable to population heterogeneity	Only applicable when MID is known To the extent that MID is uncertain, this approach will be less attractive	Consider as complementing other approaches, particularly the presentation of relative and absolute effects

Abbreviations: SD, standard deviation; MID, minimally important difference.

Table 9: Application of approaches to dexamethasone for pain after laparoscopic cholecystectomy example[43]

Outcomes	Estimated risk or estimated score/value	Absolute reduction in risk or reduction in score/value with dexamethasone	Relative effect (95% CI)	Number of participants (studies)	Confidence in effect estimate	Comments
(A) Postoperative pain, SD units: investigators measured pain using different instruments. Lower scores mean less pain	The pain score in the dexamethasone groups was on average **0.79 SDs (1.41–0.17) lower** than in the placebo groups		-	539 (5)	Low evidence[a,b]	As a rule of thumb, 0.2 SD represents a small difference, 0.5 a moderate, and 0.8 a large
(B) Postoperative pain, natural units: measured on a scale from 0 (no pain) to 100 (worst pain imaginable).	The mean postoperative pain scores with placebo ranged from 43 to 54	The mean pain scores in the intervention groups was on average **8.1 (1.8–14.5) lower**	-	539 (5)	Low evidence	Scores estimated based on an SMD of 0.79 (95% CI:1.41, 0.17). The minimally important difference on the 0e100 pain scale is approximately 10
(C) Substantial postoperative pain: investigators measured pain using different instruments	20 per 100[c]	More patients in dexamethasone group achieved important improvement in pain score **0.15 (95% CI: 0.19, 0.04)**		539 (5)	Low evidence	Scores estimated based on an SMD of 0.79 (95% CI:1.41, 0.17) Method assumes that distributions in intervention and control groups are normally distributed and variances are similar
(D) Postoperative pain: investigators measured pain using different instruments. Lower scores mean less pain	28.1[d]	3.7 lower pain score (6.1 lower 0.6 lower)		539 (5)	Low evidence	Weighted average of the mean pain score in dexamethasone group divided by mean pain score in placebo
(E) Postoperative pain: investigators measured pain using different instruments	The pain score in the dexamethasone groups was on average **0.40 (95% CI: 0.74, 0.07) minimally important difference units** less than in the control group		-	539 (5)	Low evidence	An effect less than half the minimally important difference suggests a small or very small effect

Abbreviations: CI, confidence interval; SD, standard deviation; SMD, standardized mean difference.

a. Evidence limited by heterogeneity between studies

b. Evidence limited by imprecise data

c. The 20% comes from the proportion in the control group requiring rescue analgesia

d. Crude (arithmetic) means of the postoperative pain mean responses across all five trials when transformed to a 100-point scale

Table 10 provides an example of how inconsistency is explained in an evidence profile. The footnotes highlight the large I^2 value and, while some of the heterogeneity may be explained by study limitations, there is enough concern to warrant downgrading the body of evidence. As a result, the table shows serious concerns with inconsistency.

Table 10. Evidence profile for outcome of incidence of arthralgia (0–42 days)[24]

Certainty assessment							**No. of patients**		**Effect**		**Certainty**	**Importance**
No. of studies	Study Design	Risk of Bias	Inconsistency	Indirectness	Imprecission	Other considerations	rVSV-vaccine	No rVSV-vaccine	Relative (95% CI)	Absolute (95% CI)		
6	Randomized trials	Serious[a]	Serious[b]	Not serious	Serious[c]	None	316/1874 (16.9%)	42/891 (4.7 %)	RR 2.55[d] (0.94 to 6.91)	73 more per 1,000 (from 3 fewer to 279 more)	Very Low	Critical
2	Non-randomized studies	Not serious	Not serious	Not serious	Serious[d]	None	75/469 (16.0%)	8/99 (8.1%)	RR 1.63[e] (0.0001 to 7739.16)	51 more per 1,000 (from 81 fewer to 1000 more)	Very Low	Critical

Note: Non-randomized studies without comparators are not included in evidence table, but would be considered of very low certainty (evidence type 4); CI: Confidence interval; RR: Relative risk

Explanations

a. Participants, healthcare personnel, and outcome assessors were not blinded in Huttner 2015 or Samai 2018 potentially influencing events reported for this subjective outcome. Concern for possible underreporting in Kennedy because arthralgia was only solicited at one week and at one month for most participants; Huttner only solicited arthralgia for low dose participants

b. Rated down once due to concerns with heterogeneity (I^2=70%). Some may be explained by concerns with risk of bias (poor randomization or outcome definition)

c. The 95% confidence interval of the mean pooled estimate includes potential for possible harms as well as benefits

d. Few events reported do not meet optimal information size and suggest fragility in the estimate

e. RR calculated using the standard continuity correction of 0.5 and uses a random effects model

8.3 Indirectness

Research that answers the PICO question most appropriately is considered direct evidence; therefore, studies that address the target population, compare the interventions specified in the question and measure the outcomes of interest can be classified as direct evidence[7]. Indirectness can be introduced when any of the four situations below occur:

- The population that participated in studies may differ from the population of interest;
- The intervention that was evaluated may differ from the intervention of interest;
- The primary interest is head-to-head comparisons of vaccine A to vaccine B, but A was compared with C and B was compared with C (i.e., the comparator is different from the comparator of interest)
- The outcome that was assessed may differ from that of primary interest. This may occur when there is either an intermediate outcome or a surrogate outcome used to inform the outcome of interest. For example, a panel may decide that vaccine efficacy is a critical outcome; however, the underlying evidence does not report directly on the measure of efficacy. This may occur when there is a low baseline risk of developing the outcome of interest. When assessing the evidence for vaccines, immunogenicity may serve as an appropriate surrogate for vaccine efficacy if vaccine efficacy data are not available; however, unless there is an established immune correlate of protection, this should result in downgrading for indirectness.

Table 11. Examples of indirect evidence

Indirect	**Question of Interest**	**Source of Indirectness**
Population	a. Efficacy of vaccine in preventing disease.	a. Studies are available for healthy persons, but not for the population of interest (e.g., older adults with chronic health conditions). b. Studies are available for the correct population; however, the baseline risk of infection is not representative of the recruited target population in the trial. For example, RSV vaccine trial participants are recruited during a year with unrepresentatively low RSV rates.
Intervention	Efficacy of a new formulation of a vaccine in preventing disease.	Studies of previous formulations of the vaccine provide indirect evidence bearing on the new vaccine.
Comparator	Efficacy of vaccine A compared to vaccine B in preventing disease.	Studies compared vaccine A to placebo and vaccine B to placebo, but studies comparing A to B are unavailable.
Outcome	Prevention of disease.	Increase in antibody titers following vaccination are reported, but there are no well-established standard correlates of protection.
Intervention vs. Comparator	Efficacy of vaccine A compared to no vaccine in preventing disease.	Studies only compare vaccine A to the current standard of care, vaccine B; therefore, the relationship between the intervention and the comparator is indirect.

Both systematic reviews and guidelines may require the use of evidence that is indirect with respect to the comparator and outcomes of interest. Guidelines also commonly deal with evidence that is indirectly related to the population and intervention specified in the PICO question; these are sometimes described as concerns with applicability. When limited evidence is available, it is often necessary to turn to indirect evidence to help inform judgements. For the purpose of guidelines, it is important to consider all four potential causes of indirectness when rating down the domain; when there are multiple concerns with indirectness, it may be appropriate to rate down twice for indirectness. The use of surrogate outcomes typically results in rating down unless evidence of a strong association between the surrogate and the long- or short-term outcome of interest is established. The rating down process is not always additive, thus it is important to consider the evidence from all angles.

When developing recommendations, guidelines may need to use surrogate outcomes and/or indirect evidence. Although direct evidence is ideal, recommendations may be supported by indirect evidence as long as the indirectness is acknowledged in the certainty assessment.

To decide whether JYNNEOS® (orthopoxvirus) vaccine primary series or ACAM2000 vaccine primary series should be recommended for persons who are at risk for occupational exposure to orthopoxviruses, the guideline panel prioritized the outcome of “Prevention of Disease”. However, cases of orthopoxvirus were not reported by the trials. Instead, the surrogate measures of geometric mean titer (GMT) and seroconversion rate were used to inform the outcome of “Prevention of Disease”. The work group decided to rate down for indirectness for both of these measures, as there was some uncertainty in how directly findings about the GMT or seroconversion rate would predict prevention of disease. Table 12a presents a truncated GRADE Evidence Profile showing the use of a surrogate outcome to inform the critical outcome of Prevention of Disease. The second outcome presented, Severity of Disease, was informed by one trial reporting on the proportion of study participants with an attenuated take lesion. The ideal measure of disease severity is taking maximum lesion area. However, the work group recognized that the clinical difference between categorical (proportion of participants with attenuated take) and the continuous measurement (take maximum lesion area) was minimal and therefore did not rate down for indirectness for the outcome of Severity of Disease.

In a second example, the ACIP recently provided recommendations for the following policy question: Should pre-exposure vaccination with the rVSVΔG-ZEBOV-GP vaccine be recommended for adults 18 years of age or older in the U.S. population who are at potential occupational risk of exposure to Ebola virus (species Zaire ebolavirus) for prevention of Ebola virus infection[10]. Due to the limited literature available for certain outcomes like the development of Ebola-related symptomatic illness, a randomized cluster study was used in the evidence profile that focused on contacts of recently confirmed Ebola cases in Guinea, west Africa[44]. Since the PICO question was specific to the U.S. population, the evidence was downgraded for indirectness but was still used to support the guideline recommendations. As a result, in table 12b, the cluster study is downgraded, and an explanation is provided in the footnotes regarding why there are serious concerns for indirectness.

Table 12a. GRADE Evidence Profile for Use of JYNNEOS (orthopoxvirus) vaccine primary series for research, clinical laboratory, response team, and healthcare personnel[11]

Certainty assessment							№ of patients		Effect		Certainty	Importance
№ of studies	Study design	Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations	JYNNEOS OPXV vaccine primary series	ACAM2000 OPXV vaccine primary series	Relative (95% CI)	Absolute (95% CI)		
A. Prevention of disease (assessed with: geometric mean titer)												
2[1,2,3,4,5,6]	randomized trials	not serious	not serious	serious[a,b]	not serious	none	213	199	-	MD **1.62 titer units higher** (1.32 higher to 1.99 higher)[c]	Moderate	CRITICAL
A. Prevention of disease (assessed with: seroconversion rate)												
2[1,2,3,4,5,6]	randomized trials	not serious	not serious	serious[b,d]	serious[e]	none	213/213 (100.0%)	192/199 (96.5%)	**RR 1.02** (0.99 to 1.05)	**19 more per 1,000** (from 10 fewer to 48 more)	Low	CRITICAL
B. Severity of disease (assessed with: maximum lesion area)												
1[7]	randomized trials	serious[f]	not serious	not serious[g]	very serious[e,h]	none	15/15 (100.0%)[i]	8/8 (100.0%)	**RR 1.00** (0.83 to 1.20)	**0 fewer per 1,000** (from 170 fewer to 200 more)	Very low	IMPORTANT

RR: risk ratio; **CI:** confidence interval

Explanations

a. Geometric mean titer is an indirect measure of efficacy.
b. Frey study used Dryvax in the comparison group. For the immunogenicity outcomes we do not feel there would be a significant difference between the two live vaccines.

c. In order to calculate a mean difference and 95% CI, geometric mean data were transformed to arithmetic mean. The effect estimate was then transformed to geometric mean difference, which you see here.
d. Seroconversion rate is an indirect measure of efficacy.
e. 95% CI includes the potential for both meaningful benefit as well as meaningful harm.
f. Concerns for risk of bias due to attrition. The two groups that contributed data to the intervention and comparison for this outcome lost between 11 and 21% of participants at the time this outcome was assessed.
g. The ideal measure of disease severity is to take maximum lesion area. This study reports the proportion of participants with an attenuated take lesion. Clinical difference between categorical (proportion of participants with attenuated take) vs. continuous measurement (take maximum lesion area) is minimal. We feel this won't affect indirectness. See Parrino et al. 2007 for a description of lesion attenuation criteria.

Table 12b. Evidence profile for outcome of development of Ebola-related symptomatic illness[24]

Certainty assessment							No. of patients		Effect		Certainty	Importance
No. of studies	Study Design	Risk of Bias	Inconsistency	Indirectness	Imprecission	Other considerations	rVSV-vaccine	No rVSV-vaccine	Relative (95% CI)	Absolute (95% CI)		
1	Randomized[a] (clusters)	Not serious	Not serious	Serious[b]	Serious[c]	None	0/51 (0.0%)	7/47 (14.9 %)	RR 0.06[d] (0 to 1.05)	140 fewer per 1,000 (from 149 fewer to 7 more)	Low Evidence	Critical
1	Non-randomized (participants)	Not serious	Not serious	Serious[b]	Serious[c]	Strong association	0/2108[f] (0.0%)	16/3075(0.5%)	RR 0.04[e] (0 to 0.74)	5 fewerper 1,000 (from fewer to 1 fewer)	Moderate Evidence	Critical

Note: Outcome assessed with laboratory confirmed case of EVD

Explanations

a. Henao-Restrepo 2017 was a cluster randomized trial (i.e., units of randomization were clusters); cluster-level data presented here.
b. Concern for indirectness to U.S. population: population consists of contacts and contacts of contacts of EVD case, ring vaccination strategy which may include post-exposure vaccination.
c. Because this study was done at a time when the 2014—2015 West Africa outbreak was waning in Guinea and there are few events reported, it does not meet optimal information size and suggests fragility in the estimate; 95% CI contains the potential for desirable as well as undesirable effects.
d. Henao-Restrepo 2017 was a cluster randomized trial (i.e., units of randomization were clusters); participant-level data presented here
e. The concerns with indirectness pose no inflationary effect; therefore, the evidence was rated up based on a very large magnitude of effect from the 96% reduction in risk and overall certainty was upgraded two levels.
f. Denominator represents participants from the clusters randomized to receive immediate vaccination.
g. RR calculated using the standard continuity correction of 0.5.

8.4 Imprecision

Imprecision refers to the risk of random error in the evidence. It is rated as either not serious, serious or very serious, similar to the other GRADE domains discussed above[45]. The estimated effect is considered imprecise when studies have a wide confidence interval (CI). This usually occurs when few events and few patients are included in studies. Concerns with imprecision can lead to uncertainty in the results presented in the evidence. For systematic reviews, the following indicate imprecision for an outcome:

- Total sample size across all studies for an outcome is lower than the calculated sample size for a single adequately powered study (online calculators are available for sample size calculations; or
- The 95% confidence interval (CI) of the pooled or best estimate of effect size includes both no effect AND appreciable benefit or appreciable harm (even if sample size is adequate). When an outcome is rare, 95% CIs of relative effects may be very wide, but 95% CIs of absolute effects may be narrow; in such situations, the evidence level may not be downgraded. For continuous outcomes, the threshold for appreciable benefit or appreciable harm refers to the difference in score in the outcome that is perceived as important.

For guidelines, additional considerations like clinical decision thresholds for optimal sample size and the event rate must be accounted for[46]. The evidence level may be downgraded because of imprecision in the following situations:

- When the recommendation is for an intervention, and
    - The 95% CI includes both no effect AND an effect that represent a benefit that would outweigh potential harms.
    - The 95% CI excludes no effect, but the lower confidence limit crosses a threshold below which, given potential harms, one would not recommend the intervention
- When the recommendation is against an intervention, and
    - The 95% CI includes no effect AND an effect that represent a harm that despite the benefits, would still be unacceptable.
    - The 95% CI excludes no effect, but the upper confidence limit crosses a threshold above which, given the benefits, one would recommend the intervention.

When assessing the risk for rare events (e.g., GBS, myocarditis, etc.) caused by a vaccine, the number of events needed may not be large enough to detect such rare events. The suspected rate of such events should be assessed in relation to the number of subjects tested to determine if the evidence should be downgraded for concerns about fragility with imprecision. An alternative approach would be to calculate the optimal information size (OIS) based on the total population instead of relying on the number of events that typically inform a judgment for imprecision. The OIS has been defined as the minimum amount of cumulative information required for reliable conclusions about an intervention, i.e., a calculation similar to calculating the sample size of patient in an individual trial, the difference being that the OIS considers the potential for heterogeneity between studies[47]. Therefore, if the number of participants in the meta-analyses is less than what is generated from a conventional sample-size calculation, there may be serious or very serious concerns about imprecision.

Table 11 provides an example of how imprecision assessments are justified. For example, the results from the randomized controlled trials are informed by a large sample size, however, the confidence interval is wide and cannot exclude the potential for both harm and benefit. Thus, concerns with imprecision are serious. In contrast, the results from the NRS have a wide confidence interval that cannot exclude the potential for harm and benefit; they are informed by few events that do not meet the optimal information size. Therefore, the concerns with imprecision are classified as "very serious" rather than "serious".

More information on assessing imprecision is available in the "Grade Guidelines 6. Rating the quality of evidence—imprecision" 2011[45,48]

Table 13. Evidence profile for outcome of incidence of arthritis (5-56 days)[24]

Certainty assessment							**No. of patients**		**Effect**		**Certainty**	**Importance**
No. of studies	**Study Design**	**Risk of Bias**	**Inconsistency**	**Indirectness**	**Imprecision**	**Other considerations**	**rVSV-vaccine**	**No rVSV-vaccine**	**Relative (95% CI)**	**Absolute (95% CI)**		
4	Randomized trials	Serious[a]	Not serious	Not serious	Serious[b]	None	39/1776 (2.2%)	16/868 (189 %)	RR 1.80[d] (0.21 to 15.3)	23 fewer per 1,000 (from 22 fewer to 400 more)	Low Evidence	Critical
2	Non-randomized studies	Not serious	Not serious	Not serious	Very Serious[b, d]	None	43/520 (8.3%)	3/107 (2.8 %)	RR 2.06[d] (0.0001 to 7739.16)	33 more per 1,000 (from 28 fewer to 1000 more)	Very low Evidence	Critical

Note: Non-randomized studies without comparators are not included in evidence table, but would be considered of very low certainty (evidence type 4)

Explanations

a. Studies used variable definitions and methods for diagnosing and reporting arthritis. In addition, participants, healthcare personnel, and outcome assessors were not blinded in Huttner 2015 or Samai 2018 potentially influencing events reported for this subjective outcome.
b. The 95% CI includes the potential for possible harms, as well as possible benefit.
c. Few events reported do not meet optimal information size and suggest fragility in the estimate.
d. RR calculated using the standard continuity correction of 0.5 and the overall effect uses a random effects model.

8.5 Publication bias

Publication bias is a type of reporting bias that leads to a systematic underestimation or an overestimation of the underlying effect (beneficial or harmful) due to the selective publication of studies[49]. Publication bias arises when investigators fail to publish studies, typically those that show no effect. Publication bias might be suspected if the available studies are uniformly small and funded by industry; a thorough review of clinical trial registries should be performed to identify if any trials were registered but not published. A funnel plot of studies with the magnitude of the effect size (e.g., relative risk or odds ratio for a binary outcome) on the X-axis, and variance (proxy for sample size) on the Y-axis can help assess publication bias. A funnel plot with asymmetrical distribution suggests publication bias. For meta-analyses with fewer than 10 studies, performing a funnel plot may be skewed; therefore, it is recommended to only perform when more than 10 studies are available. In situations with fewer than 10 studies, authors can consider additional factors when assessing publication bias: size and direction of identified studies, records of unpublished trials, availability of intervention under investigation (i.e., proprietary or specialty vaccines may be more regulated or documented, therefore, increased confidence that all available studies have been identified).

Due to the challenges in determining publication bias, publication bias is either described as "undetected" or "strongly suspected" in an evidence profile. Figure 7 provides an example of a funnel plot that has a symmetrical distribution and there is not suspicion of undetected publication bias. Conversely, figure 8 presents an example in which the forest plot is asymmetrical and therefore suggests there may be concerns with publication bias, requiring further investigation.

Figure 7. Example of funnel plot with no strong suspicion of publication bias[50]

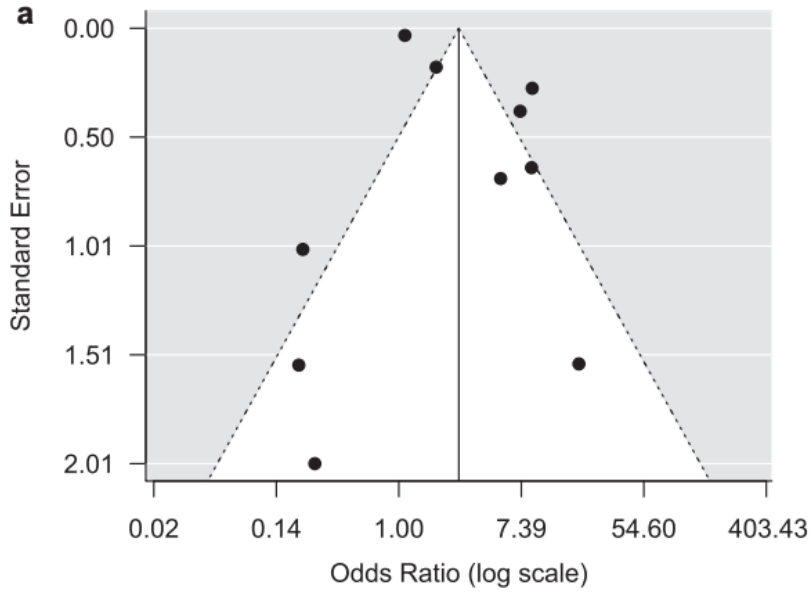
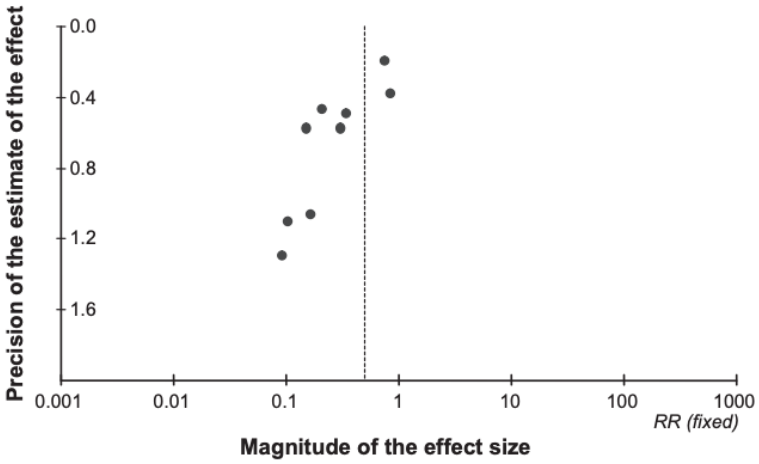


Figure 8. Example of a funnel plot with suspicion of publication bias[49]



References

7. Guyatt GH, Oxman AD, Kunz R, et al. GRADE guidelines: 8. Rating the quality of evidence--indirectness. *J Clin Epidemiol*. 2011/12// 2011;64(12):1303-1310.
10. ACIP Grading for Ebola Vaccine | CDC. 2021/01/07/T05:56:55Z 2021
24. Choi MJ, Cossaboom CM, Whitesell AN, et al. Use of ebola vaccine: recommendations of the Advisory Committee on Immunization Practices, United States, 2020. *MMWR Recommendations and Reports*. 2021;70(1):1.
31. Guyatt GH, Oxman AD, Vist G, et al. GRADE guidelines: 4. Rating the quality of evidence--study limitations (risk of bias). J Clin Epidemiol. 2011/04// 2011;64(4):407-415. doi:10.1016/j.jclinepi.2010.07.017
32. Risk of bias tools - RoB 2 tool.
33. Sterne JA, Savović J, Page MJ, et al. RoB 2: a revised tool for assessing risk of bias in randomised trials. BMJ. 2019;366
34. Higgins J, Savović J, Page M, Elbers R, Sterne J. Chapter 8: Assessing risk of bias in a randomized trial. In: Higgins J, Thomas J, Chandler J, et al, eds. Cochrane Handbook for Systematic Reviews of Interventions version 63 (updated February 2022). Cochrane; 2022. www.training.cochrane.org/handbook.
35. Sterne J, Hernán M, McAleenan A, Reeves B, Higgins J. Chapter 25: Assessing risk of bias in a non-randomized study. In: Higgins J, Thomas J, Chandler J, et al, eds. Cochrane Handbook for Systematic Reviews of Interventions version 63 (updated February 2022) Cochrane; 2022. www.training.cochrane.org/handbook.
36. GA Wells BS, D O'Connell, J Peterson, V Welch, M Losos, P Tugwell. The Newcastle-Ottawa Scale (NOS) for assessing the quality of nonrandomised studies in meta-analyses. Ottawa Hospital Research Institute. https://www.ohri.ca/programs/clinical_epidemiology/oxford.asp
37. Thomas Piggott RLM, Carlos A Cuello-Garcia, Nancy Santesso, Reem A Mustafa, Joerg J Meerpohl, Holger J Schünemann; GRADE Working Group. Grading of Recommendations Assessment, Development, and Evaluations (GRADE) notes: extremely serious, GRADE's terminology for rating down by three levels. J Clin Epidemiol. 2020;120:116-120. doi:10.1016/j.jclinepi.2019.11.019
38. (ACIP) ACoIP. Grading of Recommendations, Assessment, Development, and Evaluation

(GRADE): Use of JYNNEOS® (orthopoxvirus) vaccine heterologous for those who received ACAM2000 primary series. Centers for Disease Control and Prevention. https://www.cdc.gov/vaccines/acip/recs/grade/JYNNEOS-orthopoxvirus-heterologous.html
39. Guyatt GH, Oxman AD, Kunz R, et al. GRADE guidelines: 7. Rating the quality of evidence--inconsistency. J Clin Epidemiol. 2011/12// 2011;64(12):1294-1302. doi:10.1016/j.jclinepi.2011.03.017
40. Cynthia P Cordero ALD. Key concepts in clinical epidemiology: detecting and dealing with heterogeneity in meta-analyses. J Clin Epidemiol. 2021;130:149-151. doi:10.1016/j.jclinepi.2020.09.045
41. Gordon Guyatt YZ, Martin Mayer, Matthias Briel, Reem Mustafa, Ariel Izcovich, Monica Hultcrantz, Alfonso Iorio, Ana Carolina Alba, Farid Foroutan, Xin Sun, Holger Schunemann, Hans DeBeer, Elie A Akl, Robin Christensen, Stefan Schandelmaier. GRADE guidance 36: updates to GRADE's approach to addressing inconsistency. J Clin Epidemiol. 2023;158:70-83. doi:10.1016/j.jclinepi.2023.03.003
42. Higgins J, Li T, Deeks J. Chapter 6: Choosing effect measures and computing estimates of effect. In: Higgins J, Thomas J, Chandler J, et al, eds. Cochrane Handbook for Systematic Reviews of Interventions version 63 (updated February 2022). Cochrane; 2022. www.training.cochrane.org/handbook.
43. Guyatt GH, Thorlund K, Oxman AD, et al. GRADE guidelines: 13. Preparing summary of findings tables and evidence profiles-continuous outcomes. J Clin Epidemiol. Feb 2013;66(2):173-83. doi:10.1016/j.jclinepi.2012.08.001
44. Henao-Restrepo AM, Camacho A, Longini IM, et al. Efficacy and effectiveness of an rVSV-vectored vaccine in preventing Ebola virus disease: final results from the Guinea ring vaccination, open-label, cluster-randomised trial (Ebola Ça Suffit!). The Lancet. 2017/02/04/ 2017;389(10068):505-518. doi:10.1016/S0140-6736(16)32621-6
45. Guyatt GH, Oxman AD, Kunz R, et al. GRADE guidelines 6. Rating the quality of evidence--imprecision. J Clin Epidemiol. 2011/12// 2011;64(12):1283-1293. doi:10.1016/j.jclinepi.2011.01.012
46. Zeng L, Brignardello-Petersen R, Hultcrantz M, et al. GRADE guidelines 32: GRADE offers guidance on choosing targets of GRADE certainty of evidence ratings. J Clin Epidemiol. Sep 2021;137:163-175. doi:10.1016/j.jclinepi.2021.03.026
47. Pogue JM, & Yusuf, S. Cumulating evidence from randomized trials: utilizing sequential monitoring boundaries for cumulative meta-analysis. Controlled clinical trials. 1997;18(6):580-593.
48. Gordon H Guyatt ADO, Regina Kunz, Jan Brozek, Pablo Alonso-Coello, David Rind, P J Devereaux, Victor M Montori, Bo Freyschuss, Gunn Vist, Roman Jaeschke, John W Williams Jr, Mohammad Hassan Murad, David Sinclair, Yngve Falck-Ytter, Joerg Meerpohl, Craig Whittington, Kristian Thorlund, Jeff Andrews, Holger J Schünemann. GRADE guidelines 6. Rating the quality of evidence--imprecision. J Clin Epidemiol. 2011;64(12):1283-93. doi:10.1016/j.jclinepi.2011.01.012
49. Guyatt GH, Oxman AD, Montori V, et al. GRADE guidelines: 5. Rating the quality of evidence--publication bias. J Clin Epidemiol. 2011/12// 2011;64(12):1277-1282. doi:10.1016/j.jclinepi.2011.01.011
50. Yong PJ, Matwani S, Brace C, et al. Endometriosis and Ectopic Pregnancy: A Meta-analysis. J Minim Invasive Gynecol. 2020/02// 2020;27(2):352-361.e2.

9. Domains Increasing One's Certainty in the Evidence

After assessing study limitations, inconsistency, indirectness, imprecision and publication bias, three criteria should be considered that may warrant raising the evidence level in NRS: strength of association, dose-response gradient, and opposing plausible residual confounding or bias[51].

9.1 Strength of association

When the strength of the association is strong and the effect of the estimate is large or very large, the GRADE assessment may be upgraded due to more certainty in the results[51]. If a study has no major concerns with confounding or internal validity, then it may be appropriate to upgrade the evidence level. If the effect is large enough, the observed benefit cannot be explained by weak study design alone and instead allows for consideration that there is some confidence in the estimate of the effect. Therefore, while NRS are likely to provide an overestimate of the true effect, a strong association in the effect size may lead to stronger certainty in the evidence. The evidence level may be upgraded by one level if the relative risk from at least two studies is approximately >2 or <0.5, and it may be upgraded by two levels if the relative risk is approximately >5 or <0.2. Table 7a shows a scenario in which a NRS was upgraded due to the strong association seen in the effect size.

Table 14. Relationship between effect measure and evidence level

Strength of Association	**Effect Measure[a]**	**Evidence Level**
Strong	Relative Risk approximately >2 or <0.5 (based on consistent evidence from at least 2 studies)	Move up 1 level
Very strong	Relative Risk approximately >5 or <0.2	Move up 2 levels

[a]Relative risks of 0.5 and 0.2 correspond to vaccine efficacies of 50% and 80%, respectively. Vaccine efficacy = (1 – Relative Risk) x 100.

Table 15. Evidence profile for outcome of development of Ebola-related symptomatic illness[24]

Certainty assessment							No. of patients		Effect		Certainty	Importance
No. of studies	**Study Design**	**Risk of Bias**	**Inconsistency**	**Indirectness**	**Imprecision**	**Other considerations**	**rVSV-vaccine**	**NorVSV-vaccine**	**Relative (95% CI)**	**Absolute (95% CI)**		
1	Randomized[a] (clusters)	Not serious	Not serious	Serious[b]	Serious[c]	None	0/51 (0.0%)	7/47 (14.9 %)	RR 0.06[d] (0.94 to 6.91)	140 fewer per 1,000 (from 149 fewer to 7 more)	Low Evidence	Critical
1	Non-randomized (participants)	Not serious	Not serious	Serious[b]	Serious[c]	Strong association	0/2108[f] (0.0%)	16/3075(0.5 %)	RR 0.04[d] (0 to 0.74)	5 fewerper 1,000 (from 5 fewer to 1 fewer)	Moderate Evidence	Critical

CI: Confidence interval; RR: Relative risk; Note: Outcome assessed with laboratory confirmed case of EVD

Explanations

h. Henao-Restrepo 2017 was a cluster randomized trial (i.e., units of randomization were clusters); cluster-level data presented here.
i. Concern for indirectness to the U.S. population: population consists of contacts, and contacts of contacts of EVD cases, and ring vaccination strategy which may include post-exposure vaccination.
j. Because this study was done at a time when the 2014-2015 West Africa outbreak was waning in Guinea and there are few events reported, it does not meet optimal information size and suggests fragility in the estimate; 95% CI contains the potential for desirable as well as undesirable effects.
k. Henao-Restrepo 2017 was a cluster randomized trial (i.e., units of randomization were clusters); participant-level data presented here.
l. The concerns with indirectness pose no inflationary effect; therefore, the evidence was rated up based on a very large magnitude of effect from the 96% reduction in risk and overall certainty was upgraded two levels.
m. Denominator represents participants from the clusters randomized to receive immediate vaccination.
n. RR calculated using the standard continuity correction of 0.5.

9.2 Dose-response gradient

A dose-response gradient could upgrade the certainty of evidence assessment for NRS[51]. For example, if greater vaccine efficacy corresponds with increasing number of doses in a series, then the dose-response relationship may result in more confidence in the results. While residual confounding could contribute to the effect estimate, if the effect size is large and a dose-response gradient is observed, it is

likely that confounding alone cannot account for the strength of the association; therefore, the evidence level may be upgraded.

9.3 Opposing plausible residual confounding or bias

Both RCTs and NRSs may be impacted by plausible bias that underestimates the effect of an intervention or increases the effect when no effect was observed[51]. For example, if a vaccine is suspected of being associated with an adverse event and the publicity results in increased spontaneous reporting of the adverse event among vaccinated persons compared to that in unvaccinated persons, yet epidemiological studies find no association, the evidence level for the lack of association can be upgraded. Similarly, if an intervention is given to sicker patients and the results still show that they improved more than the control group, the actual effect is likely larger than the observed effect estimate.

References

24. Choi MJ, Cossaboom CM, Whitesell AN, et al. Use of ebola vaccine: recommendations of the Advisory Committee on Immunization Practices, United States, 2020. *MMWR Recommendations and Reports*. 2021;70(1):1.

51. Schünemann H, Higgins J, Vist G, et al. Chapter 14: Completing 'Summary of findings' tables and grading the certainty of the evidence. In: Higgins J, Thomas J, Chandler J, et al, eds. *Cochrane Handbook for Systematic Reviews of Interventions version 63 (updated February 2022)*. 2022. www.training.cochrane.org/handbook.

10. Overall Certainty of Evidence

When a systematic review is made to support recommendations, systematic review authors will rate the certainty of the body of evidence that informs each critical and important outcome[52,53]. When moving from the evidence to decision-making, one overall certainty in the evidence value (high, moderate, low, very low) is determined from all of the individual outcomes. This is informed by only the outcomes deemed critical, not the important outcomes. The overall certainty of evidence is typically made based on the critical outcome with the lowest certainty of evidence rating. For example, if the evidence profile presents the following outcomes, the overall certainty would be moderate: High certainty in a mortality reduction (Critical outcome), Moderate certainty in reduced incidence of hospitalization (Critical outcome), Low certainty in improvement in quality of life (Important outcome), Moderate certainty in increased serious adverse events (Critical outcome). This is because the lowest of the critical outcomes (mortality, hospitalization, and serious adverse events) is Moderate.

In certain situations, the overall confidence might not be based on outcomes which were pre-determined as critical for decision-making. The guideline panel may change what is considered to be critical based on the results of the systematic review. Certain positive outcomes or negative effects might have been found to occur infrequently; it is acceptable to decrease the importance of these outcomes. Similarly, the panel may identify specific critical outcomes that inform the recommendation and that may influence the overall certainty of the outcome; however, the caveat is that the certainty of the evidence cannot be higher than the critical harm.

References

52. Guyatt G, Oxman AD, 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. doi:10.1016/j.jclinepi.2012.01.006

53. Zhang Y, Coello PA, Guyatt GH, et al. GRADE guidelines: 20. Assessing the certainty of evidence in the importance of outcomes or values and preferences—inconsistency, imprecision, and other domains. *J Clin Epidemiol*. 2019/07/01/ 2019;111:83-93.

11. Communicating findings from the GRADE certainty assessment

Clear and standardized wording helps to communicate the findings from GRADE Summary of Findings or GRADE Evidence Profiles. Statements to communicate the findings are informed by the certainty of the evidence for the outcome and the size of the effect. Table 16 below provides suggested wording to convey the findings.

Table 16. Suggested narrative statements for phrasing conclusions[54]

Size of the effect estimate	**Suggested statements for conclusions** *(replace X with intervention, choose 'reduce' or 'increase' depending on the direction of the effect, replace 'outcome' with name of outcome, include 'when compared with Y' when needed)*
High certainty of the evidence	
Large effect	X results in a large reduction/increase in outcome
Moderate effect	X reduces/increases outcome X results in a reduction/increase in outcome
Small important effect	X reduces/increases outcome slightly X results in a slight reduction/increase in outcome
Trivial, small unimportant effect or no effect	X results in little to no difference in outcome X does not reduce/increase outcome
Moderate certainty of the evidence	
Large effect	X likely results in a large reduction/increase in outcome X probably results in a large reduction/increase in outcome
Moderate effect	X likely reduces/increases outcome X probably reduces/increases outcome X likely results in a reduction/increase in outcome X probably results in a reduction/increase in outcome

Small important effect	X probably reduces/increases outcome slightly X likely reduces/increases outcome slightly X probably results in a slight reduction/increase in outcome X likely results in a slight reduction/increase in outcome
Trivial, small unimportant effect or no effect	X likely results in little to no difference in outcome X probably results in little to no difference in outcome X likely does not reduce/increase outcome X probably does not reduce/increase outcome
Low certainty of the evidence	
Large effect	X may result in a large reduction/increase in outcome The evidence suggests X results in a large reduction/increase in outcome
Moderate effect	X may reduce/increase outcome The evidence suggests X reduces/increases outcome X may result in a reduction/increase in outcome The evidence suggests X results in a reduction/increase in outcome
Small important effect	X may reduce/increase outcome slightly The evidence suggests X reduces/increases outcome slightly X may result in a slight reduction/increase in outcome The evidence suggests X results in a slight reduction/increase in outcome
Trivial, small unimportant effect or no effect	X may result in little to no difference in outcome The evidence suggests that X results in little to no difference in outcome X may not reduce/increase outcome The evidence suggests that X does not reduce/increase outcome
Very low certainty of the evidence	
Any effect	The evidence is very uncertain about the effect of X on outcome X may reduce/increase/have little to no effect on outcome, but the evidence is very uncertain

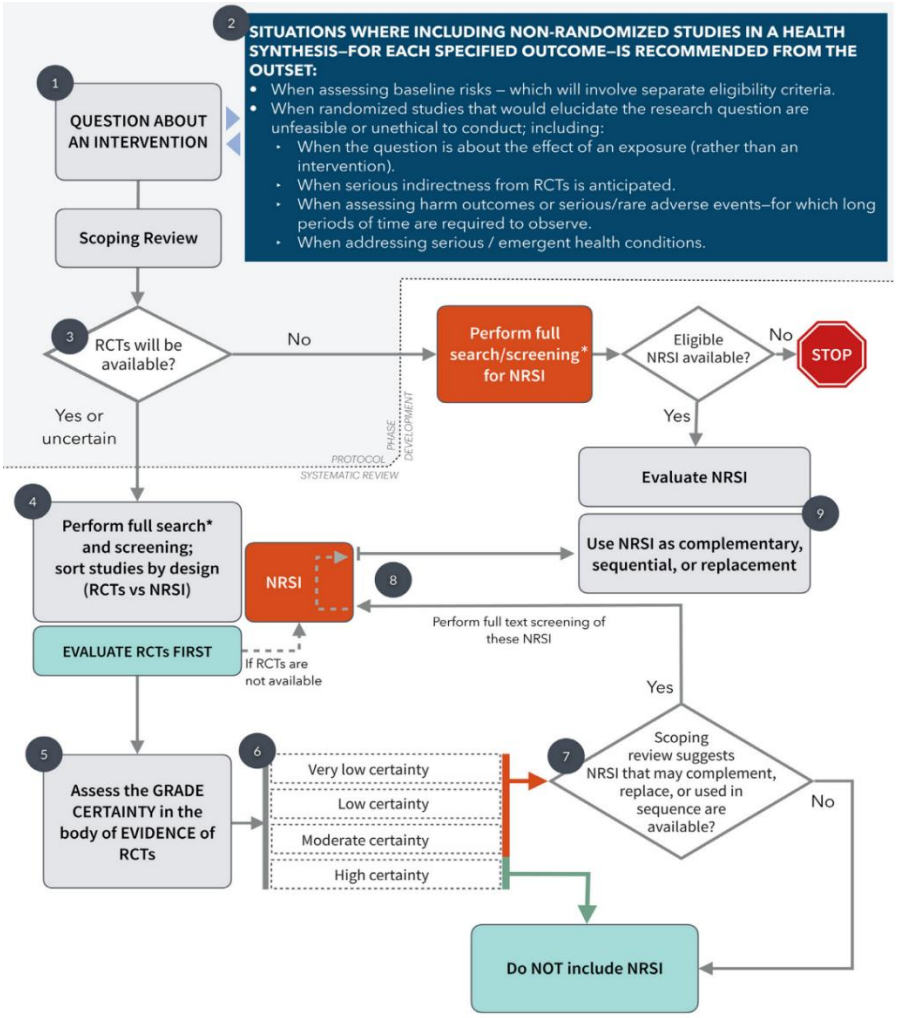
References

54. Holger J Schünemann GEV, Julian PT Higgins, Nancy Santesso, Jonathan J Deeks, Paul Glasziou, Elie A Akl, Gordon H Guyatt. Chapter 15: Interpreting results and drawing conclusions. *Cochrane Handbook for Systematic Reviews of Interventions* 2023;

12. Integrating Randomized and Non-randomized Studies in Evidence Synthesis

As described in section 4, authors at the protocol stage may decide that both RCTs and NRS need to be considered, and both types of evidence are retrieved and evaluated. Once the search is complete, the evidence is organized by study design as either randomized or non-randomized. The GRADE certainty of the RCTs should be evaluated first. After assessing each outcome separately, if there is high certainty in the body of evidence coming from RCTs, there is no need to further evaluate or use the NRS to complement or replace the RCTs. If the certainty of evidence from NRS is higher than RCTs, they can be considered as replacement evidence, especially if the NRS have low concerns with indirectness and imprecision. Reviewers might consider using NRS to complement evidence if RCTs do not provide data on populations of interest, or if the NRS studies provide evidence for possible effect modification. Figure 9 provides a visual representation of when NRS may be needed to support evidence from RCTs.

Figure 9: Flow chart depicting when to integrate RCTs and NRS in the evidence synthesis[18]



When high certainty evidence for an outcome is not available in the RCT body of evidence, NRS can be used. There are two scenarios in which this may occur[18]:

- When evidence from RCTs has low or very low certainty, NRS could help increase the overall certainty in the results. The NRS should be evaluated and if the certainty in the evidence is equal to or better than the certainty level of the RCTs, both types of evidence can be used in the decision-making process.
- When evidence from RCTs is moderate, NRS can be used to mitigate concerns with indirectness (e.g., baseline risk in the population may not represent the target population). In this situation, it is unlikely to find NRS that will have an equal or better certainty level than the RCTs as NRS can only be deemed moderate or high certainty if there is a reason to upgrade the evidence level (see section 7). It is important to remember that while NRS can provide context to RCTs, they should not be used as evidence to make judgements about directness when grading the certainty in the RCTs; directness should still be judged based on how closely the evidence answers the research question[19]. Below are two examples of how NRS help contextualize RCTs.
    - If an RCT was conducted in men and the target population in the research question was women, NRS may be used to make judgements about the certainty in these results. If the NRS shows the intervention has the same effect in both men and women, then the NRSs can be used to complement the RCT. Conversely, if the studies had shown that there was a notable difference in men and women, the overall certainty in the RCT evidence may need to be downgraded.
    - When the RCT evidence does not provide enough information about baseline risk of the control event, NRS may be used. For example, if the PICO question specified children between the ages of 12 and 15 as the target population, however the RCT evidence only provided baseline risk for children under the age of 5, NRS could be used to provide the control event rate for the target age group. The NRS could provide evidence that shows the baseline risk varies between populations or supports the evidence from the RCT.

When either of the two scenarios that result in the use of NRS occur, there are three ways in which the evidence can interact with the RCTs (Figure 10)[19]:

- Complementary NRS: The NRS can provide information on whether the intervention works similarly in different populations or if there are differential baseline risks between populations. Therefore, when the RCT evidence is indirect, NRS can be used to complement and contextualize as seen in the examples above.
- Sequential NRS: When evidence from RCTs is not sufficient, NRS can help by providing additional information. For example, NRS could provide information on long-term outcomes for patients involved in short-term RCTs. Additionally, when RCTs use surrogate outcomes, the NRS could help determine if the surrogate is relevant to patient-important outcomes.
- Replacement NRS: When the NRS is assessed and the results have a higher level of certainty than the body of evidence from RCTs, the NRS may replace the RCTs. In spite of the lack of randomization, if the NRS is more direct and has better certainty, then decision-makers can consider the NRS as the best available evidence.

Figure 10. Steps that systematic review authors might follow when considering NRS evidence (adapted) [19]

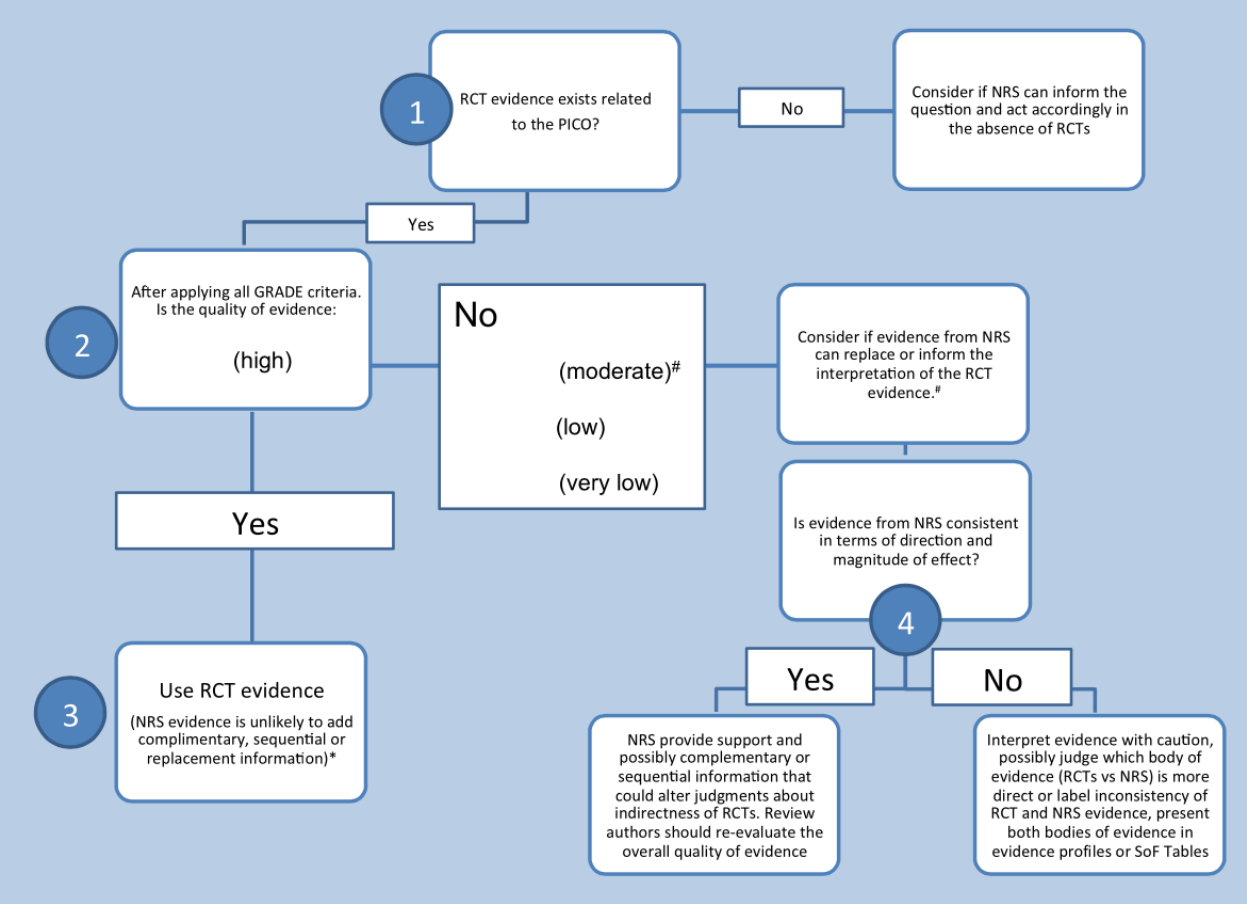


Figure 10 provides an overview of the steps taken when deciding whether to use NRS in addition to evidence from RCTs. When presenting both NRS and RCTs for an outcome in a systematic review, the results can either be presented separately as a narrative synthesis, in separate meta-analysis as a quantitative synthesis or a combination of the two[18].

References

18. Cuello-Garcia CA, Santesso N, Morgan RL, et al. GRADE guidance 24 optimizing the integration of randomized and non-randomized studies of interventions in evidence syntheses and health guidelines. *J Clin Epidemiol*. 2022/02// 2022;142:200-208. doi:10.1016/j.jclinepi.2021.11.026

19. Schünemann HJ, Tugwell P, Reeves BC, et al. Non-randomized studies as a source of complementary, sequential or replacement evidence for randomized controlled trials in systematic reviews on the effects of interventions. *Research Synthesis Methods*. 2013 2013;4(1):49-62. doi:10.1002/jrsm.1078

References

1. Ahmed F, Temte JL, Campos-Outcalt D, Schünemann HJ, Group AEBRW. Methods for developing evidence-based recommendations by the Advisory Committee on Immunization Practices (ACIP) of the US Centers for Disease Control and Prevention (CDC). *Vaccine*. 2011;29(49):9171-9176.
2. Committee on Standards for Developing Trustworthy Clinical Practice Guidelines BoHCS, Institute of Medicine. *Clinical Practice Guidelines We Can Trust*. National Academies Press; 2011.
3. Schünemann HJ, Wiercioch W, Etxeandia I, et al. Guidelines 2.0: systematic development of a comprehensive checklist for a successful guideline enterprise. *CMAJ*. 2014/02/18/ 2014;186(3):E123-E142. doi:10.1503/cmaj.131237
4. World Health O. *WHO handbook for guideline development*. World Health Organization; 2014:167.
5. Thomas J, Kneale D, McKenzie J, Brennan S, Bhaumik S. Chapter 2: Determining the scope of the review and the questions it will address. In: Higgins J, Thomas J, Chandler J, et al, eds. *Cochrane Handbook for Systematic Reviews of Interventions version 63 (updated February 2022)*. Cochrane; 2022. www.training.cochrane.org/handbook.
6. Guyatt GH, Oxman AD, Kunz R, et al. GRADE guidelines: 2. Framing the question and deciding on important outcomes. *J Clin Epidemiol*. 2011/04// 2011;64(4):395-400. doi:10.1016/j.jclinepi.2010.09.012
7. Guyatt GH, Oxman AD, Kunz R, et al. GRADE guidelines: 8. Rating the quality of evidence--indirectness. *J Clin Epidemiol*. 2011/12// 2011;64(12):1303-1310. doi:10.1016/j.jclinepi.2011.04.014
8. Fitch K, Bernstein SJ, Aguilar MD, et al. *The RAND/UCLA Appropriateness Method User's Manual*. 2001. 2001/01/01/. Accessed 2022/03/06/21:27:33. https://www.rand.org/pubs/monograph_reports/MR1269.html
9. (ACIP) ACoIP. GRADE: Use of Smallpox Vaccine in Laboratory and Health-Care Personnel at Risk for Occupational Exposure to Orthopoxviruses. Centers for Disease Control and Prevention.
10. ACIP Grading for Ebola Vaccine | CDC. 2021/01/07/T05:56:55Z 2021;
11. (ACIP) ACoIP. Grading of Recommendations, Assessment, Development, and Evaluation (GRADE): Use of JYNNEOS (orthopoxvirus) vaccine primary series for research, clinical laboratory, response team, and healthcare personnel (Policy Questions 1 and 2). Centers for Disease Control and Prevention. 2024.
12. Lefebvre C, Glanville J, Briscoe S, et al. Chapter 4: Searching for and selecting studies. In: Higgins J, Thomas J, Chandler J, et al, eds. *Cochrane Handbook for Systematic Reviews of Interventions version 63 (updated February 2022)*. Cochrane; 2022. www.training.cochrane.org/handbook.
13. Shea BJ, Reeves BC, Wells G, et al. AMSTAR 2: a critical appraisal tool for systematic reviews that include randomised or non-randomised studies of healthcare interventions, or both. *BMJ*. 2017/09/21/ 2017:j4008. doi:10.1136/bmj.j4008
14. Bristol Uo. ROBIS tool.
15. Lasserson T, Thomas J, Higgins J. Chapter 1: Starting a review. In: Higgins J, Thomas J, Chandler J, et al, eds. *Cochrane Handbook for Systematic Reviews of Interventions version 63*. 2022. www.training.cochrane.org/handbook
16. Moher D, Shamseer L, Clarke M, et al. Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. *Syst Rev*. Jan 1 2015;4:1. doi:10.1186/2046-4053-4-1
17. PROSPERO. York.ac.uk. https://www.crd.york.ac.uk/PROSPERO/

18. Cuello-Garcia CA, Santesso N, Morgan RL, et al. GRADE guidance 24 optimizing the integration of randomized and non-randomized studies of interventions in evidence syntheses and health guidelines. *J Clin Epidemiol*. 2022/02// 2022;142:200-208. doi:10.1016/j.jclinepi.2021.11.026
19. Schünemann HJ, Tugwell P, Reeves BC, et al. Non-randomized studies as a source of complementary, sequential or replacement evidence for randomized controlled trials in systematic reviews on the effects of interventions. *Research Synthesis Methods*. 2013 2013;4(1):49-62. doi:10.1002/jrsm.1078
20. DistillerSR | Systematic Review and Literature Review Software. *DistillerSR*.
21. Rayyan – Intelligent Systematic Review. https://www.rayyan.ai/
22. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*. 2021;372:n71. doi:10.1136/bmj.n71
23. Deeks J, Higgins J, Altman D. Chapter 10: Analysing data and undertaking meta-analyses. In: Higgins J, Thomas J, Chandler J, et al, eds. *Cochrane Handbook for Systematic Reviews of Interventions version 63 (updated February 2022)*. Cochrane; 2022. www.training.cochrane.org/handbook.
24. Choi MJ, Cossaboom CM, Whitesell AN, et al. Use of ebola vaccine: recommendations of the Advisory Committee on Immunization Practices, United States, 2020. *MMWR Recommendations and Reports*. 2021;70(1):1.
25. Borenstein M, Hedges LV, Higgins JP, Rothstein HR. *Introduction to meta-analysis*. John Wiley & Sons; 2021.
26. Michael Borenstein LVH, Julian P.T. Higgins, and Hannah R. Rothstein. A basic introduction to fixed-effect and random-effects models for meta-analysis. *Research Synthesis Methods*. 2010;1:97-111. doi:DOI: 10.1002/jrsm.12
27. Schünemann HJ, Cuello C, Akl EA, et al. GRADE guidelines: 18. 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*. 2019/07// 2019;111:105-114. doi:10.1016/j.jclinepi.2018.01.012
28. Morgan RL, Thayer KA, Bero L, et al. GRADE: Assessing the quality of evidence in environmental and occupational health. *Environ Int*. 2016/08//Jul- undefined 2016;92-93:611-616. doi:10.1016/j.envint.2016.01.004
29. Schünemann HJ. Interpreting GRADE's levels of certainty or quality of the evidence: GRADE for statisticians, considering review information size or less emphasis on imprecision? *J Clin Epidemiol*. 2016/07// 2016;75:6-15. doi:10.1016/j.jclinepi.2016.03.018
30. ACIP Evidence to Recommendation User's Guide (Centers for Disease Control and Prevention) (2020).
31. Guyatt GH, Oxman AD, Vist G, et al. GRADE guidelines: 4. Rating the quality of evidence--study limitations (risk of bias). *J Clin Epidemiol*. 2011/04// 2011;64(4):407-415. doi:10.1016/j.jclinepi.2010.07.017
32. Risk of bias tools - RoB 2 tool.
33. Sterne JA, Savović J, Page MJ, et al. RoB 2: a revised tool for assessing risk of bias in randomised trials. *BMJ*. 2019;366
34. Higgins J, Savović J, Page M, Elbers R, Sterne J. Chapter 8: Assessing risk of bias in a randomized trial. In: Higgins J, Thomas J, Chandler J, et al, eds. *Cochrane Handbook for Systematic Reviews of Interventions version 63 (updated February 2022)*. Cochrane; 2022. www.training.cochrane.org/handbook.
35. Sterne J, Hernán M, McAleenan A, Reeves B, Higgins J. Chapter 25: Assessing risk of bias in a non-randomized study. In: Higgins J, Thomas J, Chandler J, et al, eds. *Cochrane Handbook for Systematic Reviews of Interventions version 63 (updated February 2022)* Cochrane; 2022. www.training.cochrane.org/handbook.

36. GA Wells BS, D O'Connell, J Peterson, V Welch, M Losos, P Tugwell. The Newcastle-Ottawa Scale (NOS) for assessing the quality of nonrandomised studies in meta-analyses. Ottawa Hospital Research Institute. https://www.ohri.ca/programs/clinical_epidemiology/oxford.asp
37. Thomas Piggott RLM, Carlos A Cuello-Garcia, Nancy Santesso, Reem A Mustafa, Joerg J Meerpohl, Holger J Schünemann; GRADE Working Group. Grading of Recommendations Assessment, Development, and Evaluations (GRADE) notes: extremely serious, GRADE's terminology for rating down by three levels. *J Clin Epidemiol*. 2020;120:116-120. doi:10.1016/j.jclinepi.2019.11.019
38. (ACIP) ACoIP. Grading of Recommendations, Assessment, Development, and Evaluation (GRADE): Use of JYNNEOS® (orthopoxvirus) vaccine heterologous for those who received ACAM2000 primary series. Centers for Disease Control and Prevention. https://www.cdc.gov/vaccines/acip/recs/grade/JYNNEOS-orthopoxvirus-heterologous.html
39. Guyatt GH, Oxman AD, Kunz R, et al. GRADE guidelines: 7. Rating the quality of evidence--inconsistency. *J Clin Epidemiol*. 2011/12// 2011;64(12):1294-1302. doi:10.1016/j.jclinepi.2011.03.017
40. Cynthia P Cordero ALD. Key concepts in clinical epidemiology: detecting and dealing with heterogeneity in meta-analyses. *J Clin Epidemiol*. 2021;130:149-151. doi:10.1016/j.jclinepi.2020.09.045
41. Gordon Guyatt YZ, Martin Mayer, Matthias Briel, Reem Mustafa, Ariel Izcovich, Monica Hultcrantz, Alfonso Iorio, Ana Carolina Alba, Farid Foroutan, Xin Sun, Holger Schunemann, Hans DeBeer, Elie A Akl, Robin Christensen, Stefan Schandelmaier. GRADE guidance 36: updates to GRADE's approach to addressing inconsistency. *J Clin Epidemiol*. 2023;158:70-83. doi:10.1016/j.jclinepi.2023.03.003
42. Higgins J, Li T, Deeks J. Chapter 6: Choosing effect measures and computing estimates of effect. In: Higgins J, Thomas J, Chandler J, et al, eds. *Cochrane Handbook for Systematic Reviews of Interventions version 63 (updated February 2022)*. Cochrane; 2022. www.training.cochrane.org/handbook.
43. Guyatt GH, Thorlund K, Oxman AD, et al. GRADE guidelines: 13. Preparing summary of findings tables and evidence profiles-continuous outcomes. *J Clin Epidemiol*. Feb 2013;66(2):173-83. doi:10.1016/j.jclinepi.2012.08.001
44. Henao-Restrepo AM, Camacho A, Longini IM, et al. Efficacy and effectiveness of an rVSV-vectored vaccine in preventing Ebola virus disease: final results from the Guinea ring vaccination, open-label, cluster-randomised trial (Ebola Ça Suffit!). *The Lancet*. 2017/02/04/ 2017;389(10068):505-518. doi:10.1016/S0140-6736(16)32621-6
45. Guyatt GH, Oxman AD, Kunz R, et al. GRADE guidelines 6. Rating the quality of evidence--imprecision. *J Clin Epidemiol*. 2011/12// 2011;64(12):1283-1293. doi:10.1016/j.jclinepi.2011.01.012
46. Zeng L, Brignardello-Petersen R, Hultcrantz M, et al. GRADE guidelines 32: GRADE offers guidance on choosing targets of GRADE certainty of evidence ratings. *J Clin Epidemiol*. Sep 2021;137:163-175. doi:10.1016/j.jclinepi.2021.03.026
47. Pogue JM, & Yusuf, S. Cumulating evidence from randomized trials: utilizing sequential monitoring boundaries for cumulative meta-analysis. *Controlled clinical trials*. 1997;18(6):580-593.
48. Gordon H Guyatt ADO, Regina Kunz, Jan Brozek, Pablo Alonso-Coello, David Rind, P J Devereaux, Victor M Montori, Bo Freyschuss, Gunn Vist, Roman Jaeschke, John W Williams Jr, Mohammad Hassan Murad, David Sinclair, Yngve Falck-Ytter, Joerg Meerpohl, Craig Whittington, Kristian Thorlund, Jeff Andrews, Holger J Schünemann. GRADE guidelines 6. Rating the quality of evidence--imprecision. *J Clin Epidemiol*. 2011;64(12):1283-93. doi:10.1016/j.jclinepi.2011.01.012

49. Guyatt GH, Oxman AD, Montori V, et al. GRADE guidelines: 5. Rating the quality of evidence--publication bias. *J Clin Epidemiol*. 2011/12// 2011;64(12):1277-1282. doi:10.1016/j.jclinepi.2011.01.011
50. Yong PJ, Matwani S, Brace C, et al. Endometriosis and Ectopic Pregnancy: A Meta-analysis. *J Minim Invasive Gynecol*. 2020/02// 2020;27(2):352-361.e2. doi:10.1016/j.jmig.2019.09.778
51. Schünemann H, Higgins J, Vist G, et al. Chapter 14: Completing 'Summary of findings' tables and grading the certainty of the evidence. In: Higgins J, Thomas J, Chandler J, et al, eds. *Cochrane Handbook for Systematic Reviews of Interventions version 63 (updated February 2022)*. 2022. www.training.cochrane.org/handbook.
52. Guyatt G, Oxman AD, 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. doi:10.1016/j.jclinepi.2012.01.006
53. Zhang Y, Coello PA, Guyatt GH, et al. GRADE guidelines: 20. Assessing the certainty of evidence in the importance of outcomes or values and preferences—inconsistency, imprecision, and other domains. *J Clin Epidemiol*. 2019/07/01/ 2019;111:83-93. doi:10.1016/j.jclinepi.2018.05.011
54. Holger J Schünemann GEV, Julian PT Higgins, Nancy Santesso, Jonathan J Deeks, Paul Glasziou, Elie A Akl, Gordon H Guyatt. Chapter 15: Interpreting results and drawing conclusions. *Cochrane Handbook for Systematic Reviews of Interventions* 2023;