Title

Development of outcome-specific criteria for study evaluation in systematic reviews of epidemiology studies

Authors

Radke, Elizabeth G
Glenn, Barbara
Galizia, Audrey
et al.

Publication Date

2019-09-01

DOI

10.1016/j.envint.2019.05.078

Peer reviewed

Published in final edited form as:
Environ Int. 2019 September ; 130: 104884. doi:10.1016/j.envint.2019.05.078.

Development of outcome-specific criteria for study evaluation in systematic reviews of epidemiology studies

Elizabeth G. Radke[a,*], **Barbara Glenn**[a], **Audrey Galizia**[a], **Amanda Persad**[a], **Rebecca Nachman**[a], **Thomas Bateson**[a], **J. Michael Wright**[a], **Ana Navas-Acien**[b], **Whitney D. Arroyave**[c], **Robin C. Puett**[d], **Emily W. Harville**[e], **Anna Z. Pollack**[f], **Jane S. Burns**[g], **Courtney D. Lynch**[h], **Sharon K. Sagiv**[i], **Cheryl Stein**[j], **Glinda S. Cooper**[a,k]

[a]U.S. Environmental Protection Agency, National Center for Environmental Assessment, United States

[b]Department of Environmental Health Sciences, Columbia University Mailman School of Public Health, United States

[c]Integrated Laboratory Systems, United States

[d]Department of Epidemiology and Biostatistics, University of Maryland School of Public Health, United States

[e]Department of Epidemiology, Tulane University School of Public Health and Tropical Medicine, United States

[f]Department of Global and Community Health, College of Health and Human Services, George Mason University, United States

[g]Department of Environmental Health, Harvard T. H. Chan School of Public Health, United States

[h]Department of Obstetrics and Gynecology, The Ohio State University College of Medicine, United States

[i]Division of Epidemiology, University of California Berkeley, United States

[j]Department of Child and Adolescent Psychiatry, Hassenfeld Children's Hospital at NYU Langone, United States

[k]The Innocence Project, United States

Abstract

Introduction and objective: Systematic review tools that provide guidance on evaluating epidemiology studies are receiving increasing attention and support because their application facilitates improved quality of the review, consistency across reviewers, and transparency for readers. The U.S. Environmental Protection Agency's Integrated Risk Information System (IRIS) Program has developed an approach for systematic review of evidence of health effects from

*Corresponding author. radke-farabaugh.elizabeth@epa.gov, bethradke@gmail.com (E.G. Radke).

Supplementary data to this article can be found online at https://doi.org/10.1016/j.envint.2019.05.078.

chemical exposures that includes structured approaches for literature search and screening, study evaluation, data extraction, and evidence synthesis and integration. This approach recognizes the need for developing outcome-specific criteria for study evaluation. Because studies are assessed at the outcome level, a study could be considered high quality for one investigated outcome, and low quality for another, due to differences in the outcome measures, analytic strategies, how relevant a certain bias is to the outcome, and how the exposure measure relates to the outcome. The objective of this paper is to illustrate the need for outcome-specific criteria in study evaluation or risk of bias evaluation, describe the process we used to develop the criteria, and summarize the resulting criteria.

Methods: We used a process of expert consultation to develop several sets of outcome-specific criteria to guide study reviewers, improve consistency, and ensure consideration of critical issues specific to the outcomes. The criteria were developed using the following domains: outcome assessment, exposure measurement (specifically timing of exposure in relation to outcome; other exposure measurement issues would be addressed in exposure-specific criteria), participant selection, confounding, analysis, and sensitivity (the study's ability to detect a true effect or hazard).

Results: We discuss the application of this process to pregnancy-related outcomes (preterm birth, spontaneous abortion), other reproductive-related outcomes (male reproductive hormones, sperm parameters, time to pregnancy, pubertal development), chronic disease (diabetes, insulin resistance), and acute or episodic conditions (asthma, allergies), and provide examples of the criteria developed. For each outcome the most influential methodological considerations are highlighted including biological sample collection and quality control, sensitivity and specificity of ascertainment tools, optimal timing for recruitment into the study (e.g., preconception, specific trimesters), the etiologically relevant window for exposure assessments, and important potential confounders.

Conclusions: Outcome-specific criteria are an important part of a systematic review and will facilitate study evaluations by epidemiologists with experience in evaluating studies using systematic review methods who may not have extensive discipline-specific experience in the outcomes being reviewed.

1. Introduction

Application of systematic review methodology to environmental exposures in epidemiologic research is a developing field, with multiple tools (or adaptations of existing tools) addressing some aspect of study quality published in the last several years (Rooney et al., 2016), including Navigation Guide (Woodruff and Sutton, 2014), National Toxicology Program Office of Health Assessment and Translation (NTP, 2015a), European Food Safety Authority (EFSA, 2017), Biomonitoring, Environmental Epidemiology, and Short-Lived Chemicals instrument (Lakind et al., 2015), and the U.S. Environmental Protection Agency's Toxic Substances Control Act evaluations (U.S. EPA, 2018a). Though the specific purpose of these tools differs, they all provide guidance on how to evaluate potential bias in epidemiological studies of environmental and occupational exposures. Use of systematic review tools is intended to improve consistency across reviewers and allow them to interpret and synthesize results in the context of the reliability and validity of each study.

Regardless of the study evaluation tool being used, there is an additional need for evaluations to be exposure- and outcome-specific. A study could be considered high quality for one investigated outcome, and low quality for another, due to, for example, differences in outcome prevalence, measurement error, analytic strategies, study design, and how the timing of exposure assessment relates to the outcome definition. The identification of limitations specific to a particular outcome enhances the ability of the study evaluation tool to transparently document potential selection bias, information bias, and confounding that could distort effect estimates to varying degrees. For example, blinding of outcome assessors may be irrelevant for mortality but critical for assessing a subjective outcome like some motor function tests (Guyatt et al., 2011). Similarly, a study in which exposure is measured concurrent with the presence of a condition may be evaluated as inappropriate for investigating some outcomes, such as diabetes or cancer, when the exposure cannot reflect the relevant etiologic exposure window in earlier years. However, the same study may be acceptable when the exposure is expected to result in short-term effects, such as insulin resistance among non-diabetics. Also, within a single study, reliability of outcome ascertainment may be age-dependent (e.g., asthma in children less than five years may not be reliably ascertained); thus, an outcome could be considered adequate in one age group and critically deficient in another. This type of consideration is not included in tools such as those listed above, because they appropriately are focused on issues that apply regardless of the specific exposure and outcome being reviewed. Thus, there is a need for an additional step in the study evaluation process: the development of specific criteria to supplement the higher-level tool.

In the guidance for the Cochrane Collaboration's Risk of Bias for Non-randomized Studies of Interventions (ROBINS-I) tool, Sterne et al. (2016) emphasize the necessity for both methodological and content expertise. During the planning stages prior to implementing study evaluations, potential issues specific to the methods and content of the studies need to be identified. These go beyond the level of detail that would be included in a well-defined PECO statement that includes explicit descriptions of outcomes to facilitate identification of relevant studies. Rather, the outcome-specific issues could include any methodological or biological factors that would influence confidence in the results of the studies during study evaluation/risk of bias evaluation. For some outcomes, the general considerations or criteria provided in the tools might prove to be adequate for performing an outcome-specific evaluation, but this determination should be made by a reviewer or reviewers with subject matter knowledge to ensure that critical issues are not missed. If subject matter experts identify outcome-specific considerations, it is helpful to develop outcome-specific evaluation criteria to guide the evaluations. The criteria can then be used by reviewers without the same level of in-depth specific subject matter knowledge, while further improving consistency and ensuring that critical issues are considered. A similar approach can be used for exposure measurement criteria. However, this step of developing exposure- and outcome-specific criteria is often not done for systematic reviews or the criteria and methods for developing them are not reported. While some higher quality systematic reviews clearly have performed or plan to perform evaluation on an outcome-specific basis and provide some criteria in their papers or protocols (e.g., (NTP, 2017; Lam et al., 2016; NTP, 2015b)), in general published systematic reviews have included criteria to evaluate outcomes inconsistently with variable

degrees of specificity. When criteria are incompletely reported, it can be challenging to understand how the study ratings were reached and to compare the results of systematic reviews involving the same outcomes.

The U.S. Environmental Protection Agency's (EPA) Integrated Risk Information System (IRIS) Program has developed an approach for systematic review of evidence of health effects from chemical exposures that includes structured approaches for literature search and screening, study evaluation, data extraction, analysis and synthesis of results (including integration of human, animal, and mechanistic evidence) (Fig. 1). This approach is described in detail in the systematic review protocol for phthalates, available as part of the systematic reviews in this special issue, and in the IRIS Handbook, available at https://hawcprd.epa.gov/assessment/100000039/ (see "attachments"). As described in those documents, the study evaluation approach for observational epidemiology studies uses the principles of the domain-based ROBINS-I tool (Sterne et al., 2016), modified in order to address the specific needs of environmental epidemiology studies (see Morgan et al., 2019 and https://www.bristol.ac.uk/population-health-sciences/centres/cresyda/barr/riskofbias/robins-e/ for information on a concurrent development effort called the ROBINS-E tool). Study evaluation, also called study appraisal or risk of bias evaluation in other tools, as used in the rest of this paper, is designed to assess both risk of bias and study sensitivity (i.e., the ability of a study to observe a true effect or hazard (Cooper et al., 2016)). However, describing the methodology of this larger tool is not the intent of this paper, which focuses more narrowly on development of outcome-specific criteria for study evaluation to improve quality of the reviews by ensuring systematic consideration of critical issues, and improving transparency and consistency across raters. The purpose of this paper is to illustrate the need for outcome-specific criteria in study evaluation or risk of bias evaluation, describe the process we used to develop criteria, and summarize the resulting criteria. While these criteria were developed with the IRIS program context in mind, the considerations and the process for developing them could be easily adapted for use with other study evaluation/risk of bias tools.

2. Methods

We identified outcomes that would benefit from outcome-specific evaluation criteria in planned upcoming systematic reviews and then solicited subject matter experts (consultants) to assist in the development of these criteria. Experts were epidemiologists with several years of training and experience with designing and conducting population-based or occupational studies, with expertise in research evaluating the outcomes they contributed to, largely identified from academic settings, with diverse institutions intentionally selected.

The experts, in collaboration with IRIS scientists, developed outcome-specific evaluation criteria for the following domains (Fig. 1), which were established as part of the study evaluation approach prior to initiation of this effort: outcome ascertainment, relevant time window of exposure (one component of larger exposure measurement domain), participant selection, confounding, analysis, and other attributes, not considered in another domain, that could affect study sensitivity. Selective reporting is another domain considered in the study evaluation tool, but this was not considered to have outcome-specific components and thus

is not discussed further in this paper. Under this study evaluation approach, each domain is assigned a rating (Fig. 1) based on outcome-specific criteria as well as criteria that applied to all outcomes. These domain ratings are then combined for an overall study confidence rating for that outcome. The overall rating is reached based on reviewer judgments about the likely impact the noted deficiencies in bias and sensitivity for each domain had on the results and is not a quantitative "score". These decisions, including outcome-specific considerations, inform the synthesis and integration of the evidence looking across studies (Fig. 1). The subject matter experts developed criteria for each classification level (i.e., good, adequate, deficient, critically deficient). Because the criteria are intended to be applicable to any exposure, they were developed with the expectation that they would be a starting place for any systematic review on that outcome but would need to be reviewed and customized for the specific exposure. Most aspects of exposure measurement and other factors that require knowledge of the exposure would be considered separately with exposure-specific criteria. The criteria could be further customized to any other needs specific to the systematic review question. Some study design and conduct issues were relevant to more than one domain (e.g., timing of exposure measures is related to how participants are selected as well as exposure measurement; control of confounding is related to the data analysis). Where there was potential blurring between domains, the criteria were designed to avoid penalizing a study for the same limitation in multiple domains, so issues were assigned to the domain considered most relevant by the experts.

We piloted this process on two related outcomes: asthma and allergies-allergic sensitization. In this phase, we focused on criteria relating to outcome definitions, ascertainment methods and participant selection used in population-based and occupational studies. Two groups of five experts (listed in acknowledgements, coordinated by author WA) discussed a series of directed questions pertaining to these issues for asthma or for allergies; written responses to questions were also provided by the experts. We used this discussion to develop the evaluation criteria for these outcomes, and to refine the expert consultation process for the next set of outcomes.

In this next phase of our evaluation development process (Fig. 2), we focused on three outcome categories, and recruited two subject matter expert consultants for each (author initials in parentheses)–pregnancy outcomes (AP, EH), reproductive effects (JB, CL), and diabetes (AN-A, RP). Two or more IRIS epidemiologists without subject matter expertise (author initials GC, AG, AP) were also included on each development team. The teams had the flexibility to develop separate criteria for related outcomes or multiple outcomes encompassed in the categories. In total, outcome-specific evaluation criteria were developed for ten outcomes: two pregnancy-related (preterm birth, spontaneous abortion), two male reproductive-related (reproductive hormones, semen parameters), two that could be related to either male or female reproductive function (fecundity/time to pregnancy and pubertal development), two diabetes-related (diabetes, insulin resistance among non-diabetics), and two immune-related (asthma, allergy).

During the process of criteria development, core questions were used to define the domains, and a set of corresponding prompting and follow-up questions served to help the teams focus on details relevant to study evaluation for each domain (Table 1). A set of questions

pertaining to each domain are central to the IRIS study evaluation approach described in the introduction. Their purpose in the tool is to focus the development of more specific evaluation criteria on sources of bias and study aspects affecting sensitivity for the exposure-outcome combination being reviewed. The teams began by evaluating a set of sample studies drawn primarily from the phthalates literature, as this is a broad, robust database with a large collection of studies examining many different outcomes using a variety of designs, and exposure-specific evaluation criteria had already been developed. The studies were identified in a comprehensive literature search of phthalate epidemiology studies, and approximately six sample studies were selected for each outcome based on representing the different available study designs (and their corresponding strengths and limitations). The discussions began with identification of aspects of an ideal study (i.e., a state-of-the-art observational study design, appropriate to the definition and timeframe for the exposure and endpoints assessed, that would be expected to be free of bias for the domain) and critical deficiencies so severe as to warrant exclusion of the study from future analyses. From there, the criteria for four different rating levels were fleshed out between the two extremes. The development used a consensus building process, with each member offering suggestions regarding content and phrasing of the criteria that were discussed as a group; each team member reviewed the edited versions and offered additional suggestions for discussion if necessary. In general, 2 or 3 rounds of edits were performed for each domain. Any discrepancies of opinion were resolved with further discussion. Once a working set of draft criteria were available, the teams applied them to a second set of sample studies and made revisions as necessary.

As part of this process, it became clear that some considerations/criteria were not outcome-specific, but rather applicable to all or most outcomes. This type of criteria was shared across teams frequently as they were developed to maintain consistency and aid efficiency. Each team could take what had been done already, provide their input and edits, and then share with the other teams. This had the advantage of establishing multi-stakeholder consensus for non-outcome-specific considerations rather than duplicating effort across the teams and producing documents that would require extensive resolution. Outcome-specific criteria were not shared across teams.

Once the criteria were drafted, a team of two to four epidemiologists (author initials GC, ER, BG, AG, AP, TB, MW, acknowledgment initials LP), including at least two IRIS epidemiologists with PhDs in epidemiology, applied the evaluation criteria to a set of 5–15 studies, depending on the number of studies available in the database (i.e., all of the studies in the database were used in testing). For each study, ratings were established for each domain and for overall study confidence (Fig. 1). Members of the team performed evaluations independently, and then discussed discrepancies in evaluation results and difficulties in applying the criteria and came to a consensus rating. When differences occurred, the discussion focused on whether the difference was due to ambiguity in the criterion that needed clarification or a mistake on the part of one of the reviewers (e.g., a reviewer not seeing information that had been provided in the article). Based on this experience of evaluation and reconciliation, the criteria were revised again as needed to address additional issues regarding clarity or intent identified during this process. These edits were generally minor language clarifications, but subject matter experts were available for consultation during the testing phase to answer any questions that arose, and a small

number of more substantial changes were returned to them for input. For example, a strict cutoff of participation rate for each rating level was revised to allow for more reviewer judgment in the likely impact of participation rate on the potential for selection bias for each specific study (i.e., considering rationale for why participation is unlikely to be related to exposure). Additional testing, including evaluation of studies from other chemicals, was performed when inconsistency in ratings remained.

3. Results

3.1. Outcome-specific considerations by domain

3.1.1. Outcome ascertainment—This domain is the one that is most unique to each set of outcome-specific criteria. Each outcome uses specific methodologies to determine a "case" and each methodology has inherent specificity, sensitivity and reliability. Criteria for the outcome domain had clearly defined requirements for measurements to be valid (e.g., fasting, sample collection in the morning, exclusion of known diabetics, questionnaires validated in culturally appropriate populations), and other information needed to determine whether the methods were carried out appropriately (e.g., laboratory assays, quality control procedures). Reference to existing guidance or consensus publications on the outcome measurement by expert professional organizations or leading institutions in the field is informative in this section.

3.1.2. Exposure measurement (relevant time window)—Most exposure concerns are addressed in separate exposure-specific criteria, which would be developed for the chemical or other exposure being reviewed. The outcome-specific criteria for the exposure domain focuses on the timing of exposure. For example, this could include: (a) whether exposure could be measured concurrent with the outcome; (b) whether the exposure must be measured prior to development of the outcome, or (c) whether the exposure must be measured during a critical window of exposure (such as gestation, where the critical window of exposure can be a specific trimester). When applying a set of outcome-specific criteria to a specific exposure, all criteria should be reviewed for relevance to a specific chemical; however, it is particularly important to review timing of exposure measures, as differences in half-life or other exposure factors may influence what timing is acceptable. For persistent chemicals, exposure measurement after a diagnosis was made may be acceptable if it is highly likely that behavioral changes after diagnosis did not affect exposure. For chemicals with short half-lives, measuring exposure after a diagnosis may not be acceptable. Outcome-specific factors such as latency and reversibility with the removal of exposure could also influence this domain.

3.1.3. Participant selection—Much of the criteria developed for participant selection applies to all outcomes, such as the reporting of inclusion/exclusion criteria, the participation rate, and comparison of participants and non-participants. There are two main types of outcome-specific considerations for this domain. One is the timing of the participant entry into the study. For example, the evaluation criteria may strongly favor study entry in the first trimester of pregnancy (for preterm birth), or require the population under study to encompass an appropriate age range (e.g., around pubertal onset for pubertal

development). Another consideration is whether inclusion/exclusion criteria are appropriate for the population under study. For example, when studying diabetes incidence in a cohort study, diabetics should be excluded at baseline; current asthma symptoms, defined as the occurrence of asthma symptoms or medication use in the last 12 months, should be ascertained among individuals with a previous asthma diagnosis; and for semen parameters, population-based selection (with high participation rates) would be preferred to selection at infertility clinics.

3.1.4. Confounding—Most of the criteria developed for the evaluation of confounding are not outcome-specific, and thus could be applied to all outcomes. The focus for evaluation of confounding is the strategy for identifying confounders. Although a list of known risk factors was developed for each outcome, the lists are not variables that are required for a specific rating (e.g., *Good* or *Adequate*), since association with the exposure is also required for confounding to bias the results, and that cannot be captured in exposure-agnostic evaluation criteria. Rather, variables on the list should be considered as possible confounders, and the strategy for consideration can involve details relating to participant selection and characteristics, study design, and analytic approach.

3.1.5. Analysis—As with confounding, most of the criteria for evaluation of analysis are not outcome-specific. The criteria include consideration of whether appropriate analysis methods are used, whether quantitative results are presented, and whether there is adequate analysis of the robustness of the findings (e.g., sensitivity analyses). In this group of outcomes, the primary outcome-specific consideration for analysis is the characterization of the outcome variable. For some outcomes, a dichotomous or three-level variable is considered optimal (e.g., preterm birth), while for others, a continuous variable is preferred. The rationale for these preferences is described in the complete sets of criteria.

3.1.6. Sensitivity—Sensitivity issues include a narrow or low exposure range, small sample size, a high percentage below the limit of detection, an inappropriate length of follow-up, and inappropriate choice of referent group (Cooper et al., 2016). Most of these considerations are not outcome-specific, but there are some that are. It is important to note that some issues that reduce sensitivity may be considered in other domains and would not be double counted in the overall study confidence evaluation. Sensitivity is not evaluated on a four-level scale like the other domains, but rather as “adequate” or “deficient”. Since many of the sensitivity considerations are continuous measures, specific criteria were not developed, and classification of this domain relies on review-specific expert judgment on the impact of these limitations.

3.2. Application to study evaluation for specific outcomes

The key considerations identified for each outcome are listed in Table 2, and each outcome has a supplemental file with the full set of criteria. An example of criteria by classification level for diabetes and insulin resistance is in Table 3 and includes criteria that apply to all outcomes identified during the process. A case study of the application of these criteria on a paper by James-Todd et al. (2012) using NHANES data is available at: https://hawcprd.epa.gov/rob/study/100501814/ and additional examples of the application of these

criteria across several studies are available in systematic reviews of phthalates (Radke et al., 2018, Radke et al., 2019a,b (forthcoming)). The following discussion highlights some of the important considerations brought out by the expert consultations for each of the major outcome categories.

3.3. Pregnancy outcomes

Pregnancy outcomes present some unique issues due to the short duration and couple dependent nature of a pregnancy. Two pregnancy outcomes were examined during this process: preterm birth (Supplement A) and spontaneous abortion (Supplement B). For both outcomes, timing of entry into the study is important. For preterm birth, participants will ideally be enrolled during the first trimester, since with later study entry some participants who experienced early preterm birth could be excluded. For spontaneous abortion, study entry is ideally preconception (e.g., couples trying to conceive) in order to completely ascertain pregnancy losses, including early (i.e., before clinical detection) loss, but must be in the first trimester. When considering outcome ascertainment, there are different considerations for the two outcomes. For preterm birth, the outcome has a clear definition (i.e., birth prior to 37 weeks gestation), but the sensitivity and specificity of this definition is reliant on the method used to ascertain gestational age. Early pregnancy dating with ultrasound is preferred to less reliable methods, such as retrospective questionnaire or use of administrative records. For spontaneous abortion, ideal measurement of early loss is through the use of daily urine samples, which enable detection of loss prior to clinical ascertainment of pregnancy. Such early detection is not possible in a study that recruits women after clinical recognition of a pregnancy. In those studies, clinical pregnancy loss, which occurs after clinical recognition of pregnancy and prior to 20 weeks of gestation, is more feasibly ascertained. For both outcomes, exposure measurement should take place prior to the outcome occurring, around conception or during pregnancy. In addition, there are outcome-specific confounding and analysis considerations, and those are available in the supplements.

3.4. Reproductive effects

The criteria for the two male reproductive outcomes (semen parameters [Supplement C] and male reproductive hormones [Supplement D]) focused primarily on the accurate measurement of biological samples (semen and blood, respectively). For semen parameters, concentration, motility, and morphology were the primary measures of interest, and criteria address collection procedures, abstinence time, time from collection to analysis, and laboratory methods. For male reproductive hormones, testosterone, luteinizing hormone, follicle-stimulating hormone, and sex-hormone binding protein were discussed. For all of these hormones, the laboratory methods and quality control were important. In addition, for testosterone, which has diurnal variation, the criteria specify that collection must be in the morning or time of collection addressed in the analysis for a study to be considered acceptable. The criteria for semen parameters also address some specific issues for participant selection, where it is preferred, but not required, that selection occur at a setting other than an infertility clinic. For both of these outcomes, it was considered acceptable to measure the exposure concurrent with the outcome due to the potential for

short term response, but it is preferred to more closely alight with the etiologic window (e.g., 90 days prior to sperm collection to account for spermatogenesis).

For the other two reproductive outcomes (time to pregnancy [Supplement E] and pubertal development [Supplement F]), timing of both exposure measurement and study enrollment was more important. For these, the ideal design is a prospective cohort with enrollment and exposure measurement prior to development of the outcome (pregnancy and puberty, respectively). These outcomes also have specific needs for outcome ascertainment. For time to pregnancy, it is preferred to have women from couples discontinuing contraception with close monitoring for pregnancy, but other methods, including retrospective recall by women, are acceptable as well. For pubertal development, Tanner stages can be used for boys and girls, ideally by a trained examiner. Testicular volume and spermarche can also be used for boys, and menarche can be used for girls. For all four reproductive outcomes, there are outcome-specific confounding and analysis considerations, available in the supplements.

3.5. Diabetes and insulin resistance

Studies of diabetes and insulin resistance were also considered (Supplement G), with each outcome presenting unique challenges. For both outcomes, timing of exposure measurement is an important consideration, particularly for exposures with short half-lives. For diabetes, it is important for studies to exclude individuals with diabetes at baseline and include only individuals with incident disease as cases, with exposure measured prior to development of diabetes to establish temporality. Accordingly, prospective designs are generally needed. For insulin resistance, the exposure and outcome can be assessed concurrently as the outcome can be a short-term response. There are also several important criteria for outcome ascertainment. To identify undiagnosed diabetes cases, use of the American Diabetes Association definition is preferred. It includes information about fasting, laboratory test requirements and assays, quality control procedures, and measure reliability/validity. Self-reported physician diagnosis or medical treatment was deemed appropriate to identify diagnosed cases, but additional testing is needed to identify undiagnosed cases. For insulin resistance, individuals with diabetes must be excluded as measures of insulin are not interpretable in the presence of diabetes, especially if diabetes is treated with hypoglycemic medication, as treatment influences insulin production and secretion. Fasting is also required for insulin and glucose measurements. The homeostatic model assessment of insulin resistance (HOMA-IR) is the preferred measure of insulin resistance and is calculated using measurements for insulin and glucose. In addition, there are outcome-specific confounding and analysis considerations, and those are available in the supplements.

3.6. Allergy and asthma

Asthma (Supplement H) and allergy (Supplement I) are related outcomes that that require different approaches depending on whether the outcome is the development of disease (incidence) or symptoms and morbidity among those with prevalent disease. For example, the key exposure window for incident asthma (as well as allergic sensitization) is up to two years prior to diagnosis, while concurrent exposure is most informative for asthma-related symptoms among those with an asthma diagnosis as well as those with allergy-related outcomes. Physician confirmed asthma diagnosis is considered best, but a self-report of

physician diagnosed asthma can be used. Self-reported allergy outcomes using a validated questionnaire is preferable to a physician's diagnosis for allergy symptoms as many adults and children self-treat these symptoms at home and do not seek medical treatment. Age should be considered an effect modifier for both asthma and allergy outcomes, with adults and children analyzed separately. This is particularly true with asthma, as adult asthma is often very different from childhood asthma. Asthma cannot accurately be studied in children under 5, and so these children should not be enrolled in studies of asthma symptoms.

4. Discussion

The development and use of outcome-specific criteria that are used to supplement general study evaluation tools for evaluating confidence in individual epidemiology studies is an important step of the systematic review process that has not been well documented, and should be documented in the systematic review protocol. A body of literature on specific health outcomes in relation to chemical exposures may span several decades and comprise a broad spectrum of methodological approaches reflecting an evolving and maturing discipline, as well as differences in reporting detail. The involvement of epidemiologists with subject matter expertise in the outcomes and exposures that are under review is critical to assure that changes in definitions and measurement protocols that have occurred over time are captured, with an understanding of any consequent impacts on the sensitivity and specificity of outcome assessments. This understanding can be translated into criteria for evaluating potential bias and sensitivity for a systematic review. Health assessments of environmental exposure to chemicals and other hazards can involve a broad set of health effects, and ensuring that reviews are conducted by epidemiologists with subject-specific training (e.g., reproductive or respiratory epidemiology) can be challenging. Developing clear criteria also improves transparency of the systematic review.

We have described the methods and results of the approach used by IRIS to develop outcome-specific criteria for a variety of conditions occurring across the lifespan, including pregnancy outcomes, reproductive effects, chronic disease, and acute (episodic) conditions. As mentioned previously, these criteria were developed for use in IRIS assessment products but could be adapted for use with other study evaluation tools. There are several important considerations to note when developing or applying outcome-specific criteria. First, epidemiologists with discipline-specific expertise (e.g., reproductive or respiratory disease) add critical insight into the development of study evaluation tools for bias and sensitivity in environmental and occupational epidemiological studies. Evaluation of studies without input from subject-matter experts would likely miss important nuances. For example, when measuring testosterone, it is important for blood collection to be in the morning, or, if that is not possible, for time of collection to be addressed in the analysis. Lack of this adjustment is considered a critical deficiency by the subject matter experts, but an epidemiologist without reproductive expertise would likely not have identified this as a source of bias. Even with relevant expertise, however, testing and revision of the criteria are needed to ensure their consistent use and interpretation.

Each set of outcome-specific criteria is not intended to stand-alone, but rather to supplement the existing study evaluation considerations (Table 1 or another tool). The

criteria are intended to facilitate reviews by trained, experienced epidemiologists who may not have extensive discipline-specific experience in the outcomes being reviewed. While the availability of these criteria may lessen the requirement of extensive subject-matter expertise, judgments are still required based on understanding of concepts and best practices integral to the field of epidemiology, which is consistent with other programs performing systematic reviews (Rooney et al., 2016; NIEHS, 2015). Conclusions about the impact of potential bias on reported effect measure estimates require gathering information from multiple aspects of a study publication including descriptions of study methods, study groups and data analyses, as well as supporting documentation provided in supplemental documents and earlier publications. This strategy and the evaluation criteria were tested using epidemiologists with doctoral-level training (or masters-level with additional experience), and we caution against generalizing this experience to people with limited knowledge of epidemiology methods and principles.

These outcome-specific criteria are not prescriptive and should not be used as a checklist or as a tool to assign numeric scores for evaluating studies. Rather, they are intended to guide the reviewer to make their own expert judgment by highlighting relevant considerations and their relative importance. Scoring tools or numeric scales frequently mix reporting quality and applicability along with internal validity, and may not accurately account for differences in the impact of a given set of possible biases (Higgins and Green, 2011; Juni et al., 1999; Greenland, 1994). Evaluation should be guided but open-ended and reliant on expert judgment. This includes the ability to consider whether an identified limitation is likely to result in a substantial bias in the effect estimate and the need to account for a limitation in the overall study rating. Limitations identified in some domains may be deemed more important than limitations in other domains, but there is not a set of weights that is applicable to all studies. Judgment is also required on the impact of desired information that is not reported in the paper, which is a frequent occurrence.

In addition, the epidemiologists performing the study evaluations should familiarize themselves with the research on the health effects or outcomes being reviewed and consider the relevance and appropriateness of each criterion to the exposure scenario(s) being assessed. In particular, the timing of exposure measurements in the available studies (e.g., concurrent with outcome assessment or at some point prior) may guide selection of the outcome definitions to be included in the review. For example, among multiple asthma outcomes assessed in a cross-sectional study of indoor air pollution using average concentrations collected concurrent with the assessment of asthma, the definition for current asthma (i.e., asthma symptoms during the past 12 months) would be most relevant to the exposure assessment paradigm rather than an ever/never lifetime diagnosis of asthma.

There are some important limitations of this work. The criteria in the supplemental files is based on the consensus opinions of two subject matter experts and is supported by current literature in the field. However, as with any effort that relies on scientific judgment, there is potential for personal bias to be interjected into the process by individual researchers (Gotzsche and Ionnidis, 2012). Thus, including additional experts would be ideal to decrease the potential for bias, though this may not often be possible due to resource and time limitations facing many federal agencies, regulatory, authorities, and others

performing systematic reviews. There are also other sources of curated outcome-specific information developed for clinical research that were not included in this effort, such as core outcome sets (Clarke and Williamson, 2016) and the PhenX Toolkit (Hamilton et al., 2011). These resources are not designed in the context of systematic review and study evaluation and do not include all the outcome measures that are included in epidemiology studies. Nonetheless, have already undergone review and could be useful supplements or resources when developing outcome-specific criteria in the future. Another limitation is that development and testing of these criteria was based on a small selection of studies, and thus the criteria do not represent all the possible scenarios that may be evaluated in the future. Rather, it is expected that the criteria will be "evergreen" and continue to develop and evolve as they are used. It is simply not practical to foresee every outcome-specific eventuality when developing an initial set of criteria, and this combined with the evolving nature of science and the need to include exposure-specific considerations, means that the versions of the criteria in the supplement cannot be used "off the shelf" in a new systematic review, and will need to be modified and tested for the specific review need. There is no expectation that the criteria will ever truly be "final".

We have described a process for developing outcome-specific criteria for evaluation of individual epidemiology studies. Such criteria development is an important step in the systematic review process and contributes to transparency and consistency. Development of these outcome-specific criteria was a large undertaking that required contributions from many epidemiologists. However, less extensive outcome-specific criteria may be adequate, depending on the health outcome being evaluated. Nevertheless, it is a worthwhile and important step to have someone with research expertise in the outcome under consideration to review the study evaluation criteria to improve transparency and ensure that key issues are addressed. This should be an explicit and documented part of future systematic reviews.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

We would like to acknowledge the following individuals for their assistance with this project: Asthma and Allergy subject matter experts: Hasan Arshad, Peter Gergen, Elizabeth Matsui, Dan Norback, Matthew Perzanowski, Lara Akinbami, Christine Joseph, Felicia Rabito, Carl-Gustaf Bornehag; ICF contractors supporting this project: Michelle Cawley, Ashley Williams; IRIS management: Susan Rieth, Kris Thayer; External epidemiologists providing input during criteria testing: John Meeker, Wendi Robins, Maria Grau Perez; Other IRIS scientific support, including piloting of the criteria: Larissa Pardo, Leonid Kopylev; IRIS student support: Carolyn Gigot, Swati Gummadi, Alex Horansky, Jeff Nathan; EPA technical reviewers: Kristen Rappazzo and Ellen Kirrane.

References

Clarke M, Williamson P, 2016. Core outcome sets and systematic reviews. Syst. Rev 5, 11. [PubMed: 26792080]

Cooper G, Lunn R, Agerstrand M, Glenn B, Kraft A, Luke A, Ratcliffe J, 2016. Study sensitivity: evaluating the ability to detect effects in systematic reviews of chemical exposures. Environ. Int 92–93, 605–610. 10.1016/j.envint.2016.03.017.

EFSA (European Food Safety Authority), 2017. Guidance on the use of the weight of evidence approach in scientific assessments. EFSA J. 15, 4971. 10.2903/j.efsa.2017.4971.

Gotzsche P, Ionnidis P, 2012. Content area experts as authors: helpful or harmful for systematic reviews and meta-analyses. BMJ 345.

Greenland S, 1994. Quality scores are useless and potentially misleading: reply to re: a critical look at some popular analytic methods [editorial]. Am. J. Epidemiol 140, 300–301. 10.1093/oxfordjournals.aje.a117250.

Guyatt GH, Oxman AD, Vist G, Kunz R, Brozek J, Alonso-Coello P, Montori V, Akl EA, Djulbehovic B, Falck-Ytter Y, Norris SL, Williams JW, Atkins D, Meerpohl J, Schunemann HJ, 2011. GRADE guidelines: 4. Rating the quality of evidence-study limitations (risk of bias). J. Clin. Epidemiol 64, 407–415. 10.1016/j.jclinepi.2010.07.017. [PubMed: 21247734]

Hamilton C, Strader L, Pratt J, Maiese D, Hendershot T, Kwok R, Hammond J, Huggins W, Jackman D, Pan H, Nettles D, Beaty T, Farrer L, Kraft P, Marazita M, Ordovas J, Pato C, Spitz M, Wagener D, Williams M, Junkins H, Harlan W, Ramos E, Haines J, 2011. The PhenX toolkit: get the most from your measures. Am. J. Epidemiol 174, 253–260. [PubMed: 21749974]

Higgins J, Green S, 2011. Cochrane Handbook for Systematic Reviews of Interventions. Version 5.1.0 The Cochrane Collaboration 2011. http://handbook.cochrane.org.

James-Todd T, et al. , 2012. Urinary phthalate metabolite concentrations and diabetes among women in the National Health and Nutrition Examination Survey (NHANES) 2001–2008. Environ. Health Perspect 120, 1307–1313. [PubMed: 22796563]

Juni P, Witschi A, Bloch R, Egger M, 1999. The hazards of scoring the quality of clinical trials for meta-analysis. JAMA 282, 1054–1060. [PubMed: 10493204]

Lakind JS, Goodman M, Barr DB, Weisel CP, Schoeters G, 2015. Lessons learned from the application of BEES-C: systematic assessment of study quality of epidemiologic research on BPA, neurodevelopment, and respiratory health. Environ. Int 80, 41–71. 10.1016/j.envint.2015.03.015. [PubMed: 25884849]

Lam J, Sutton P, Kalkbrenner A, Windham G, Halladay A, Koustas E, Lawler C, Davidson L, Daniels N, Newschaffer C, Woodruff T, 2016. A systematic review and meta-analysis of multiple airborne pollutants and autism spectrum disorder. PLoS One 11, e0161851. 10.1371/journal.pone.0161851. [PubMed: 27653281]

Morgan R, Thayer K, Santesso N, Holloway A, Blain R, Eftim S, Goldstone A, Ross P, Ansari M, Ekl E, Filippini T, Hansell A, Meerpohl J, Mustafa R, Verbeek J, Vinceti M, Whaley P, Schunemann H, GRADE Working Group, 2019. A risk of bias instrument for non-randomized studies of exposures: a users' guide to its application in the context of GRADE. Environ. Int 122, 168–184. [PubMed: 30473382]

NIEHS (National Institute for Environmental Health Sciences), 2015. Handbook for Preparing Report on Carcinogens Monographs. U.S. Department of Health and Human Services, Office of the Report on Carcinogens, https://ntp.niehs.nih.gov/ntp/roc/handbook/roc_handbook_508.pdf.

NTP (National Toxicology Program), 2015a. Handbook for conducting a literature-based health assessment using OHAT approach for systematic review and evidence integration. U.S. Dept. of Health and Human Services, National Toxicology Program https://ntp.niehs.nih.gov/ntp/ohat/pubs/handbookjan2015_508.pdf.

NTP (National Toxicology Program), 2015b. NTP Monograph: Identifying Research Needs for Assessing Safe Use of High Intakes of Folic Acid [NTP]. National Institute of Environmental Health Sciences, Research Triangle Park, NC. http://ntp.niehs.nih.gov/ntp/ohat/folicacid/final_monograph_508.pdf.

NTP (National Toxicology Program), 2017. Report on carcinogens protocol: Methods for preparing the draft report on carcinogens monograph on antimony trioxide and other antimony compounds. In: Running Title - Antimony: RoC Protocol. National Institute of Environmental Health Sciences, Research Triangle Park, NC.

Radke EG, Braun JM, Meeker JD, Cooper GS, 2018. Phthalate exposure and male reproductive outcomes: a systematic review of the human epidemiological evidence. Environ. Int 121 (Part 1), 764–793. 10.1016/j.envint.2018.07.029. [PubMed: 30336412]

Radke EG, Galizia A, Thayer KA, Cooper GS, 2019a. Phthalate exposure and metabolic effects: a systematic review of the human epidemiological evidence. Environ. Int, 104768. 10.1016/j.envint.2019.04.040. [PubMed: 31196577]

Radke EG, Galizia A, Thayer KA, Cooper GS, 2019b. Phthalate exposure and metabolic effects: a systematic review of the human epidemiological evidence. Environ. Int, 104768. 10.1016/j.envint.2019.04.040. [PubMed: 31196577]

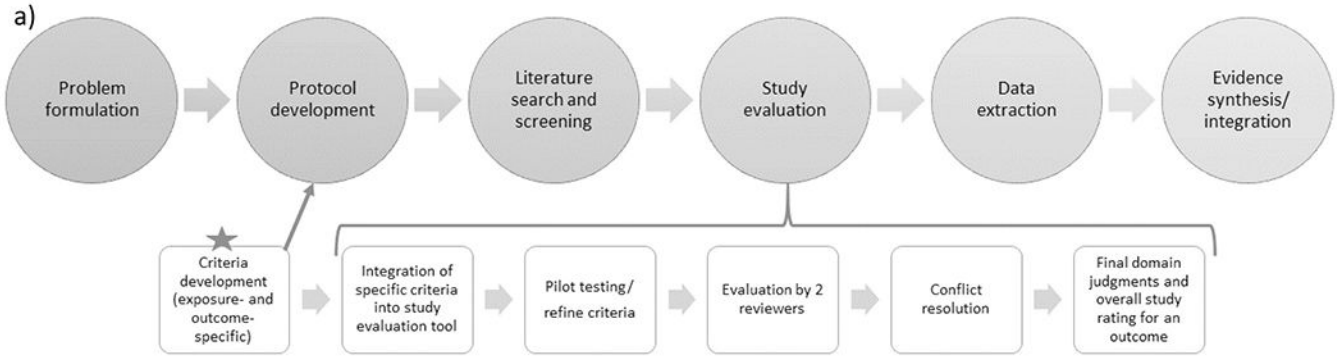
Rooney AA, Cooper GS, Jahnke GD, Lam J, Morgan RL, Boyles AL, Ratcliffe JM, Kraft AD, Schünemann HJ, Sehwingl P, Walker TD, Thayer KA, Lunn RM, 2016. How credible are the study results? Evaluating and applying internal validity tools to literature-based assessments of environmental health hazards. Environ. Int 92–93, 617–629. 10.1016/j.envint.2016.01.005.

Sterne J, Higgins J, Reeves B, 2016. ROBINS-I: a tool for assessing risk of bias in non-randomized studies of interventions, version 7 march 2016. Br. Med. J 355, i4919. [PubMed: 27733354]

U.S. EPA (U.S. Environmental Protection Agency), 2018a. Application of systematic review in TSCA risk evaluations. (740-P1-8001). U.S. Environmental Protection Agency, Office of Chemical Safety and Pollution Prevention, Washington, D.C. https://www.epa.gov/sites/production/files/2018-06/documents/final_application_of_sr_in_tsca_05-31-18.pdf.

Woodruff TJ, Sutton P, 2014. The navigation guide systematic review methodology: a rigorous and transparent method for translating environmental health science into better health outcomes [review]. Environ. Health Perspect. 122, 1007–1014. 10.1289/ehp.1307175. [PubMed: 24968373]

a)



b)

Evaluation domains
Exposure measurement
Outcome ascertainment
Participant selection
Confounding
Analysis
Sensitivity
Selective reporting

Domain judgment	Interpretation
Good	Appropriate study conduct relating to the domain and minor deficiencies not expected to influence results.
Adequate	A study that may have some limitations relating to the domain, but they are not likely to be severe or to have a notable impact on results.
Deficient	Identified biases or deficiencies interpreted as likely to have had a notable impact on the results or prevent reliable interpretation of study findings.
Critically Deficient	A serious flaw identified that makes the observed effect(s) uninterpretable. Studies with a critical deficiency will almost always be considered "uninformative" overall.

Overall Study Rating	Interpretation
High	No notable deficiencies or concerns identified; potential for bias unlikely or minimal; sensitive methodology.
Medium	Possible deficiencies or concerns noted, but resulting bias or lack of sensitivity is unlikely to be of a notable degree.
Low	Deficiencies or concerns were noted, and the potential for substantive bias or inadequate sensitivity could have a significant impact on the study results or their interpretation.
Uninformative	Serious flaw(s) makes study results unusable for hazard identification or dose response.

Fig. 1.

a) Illustration of how outcome-specific criteria fits into systematic review process; b) Study evaluation domains and evaluation classifications for determining overall study confidence.

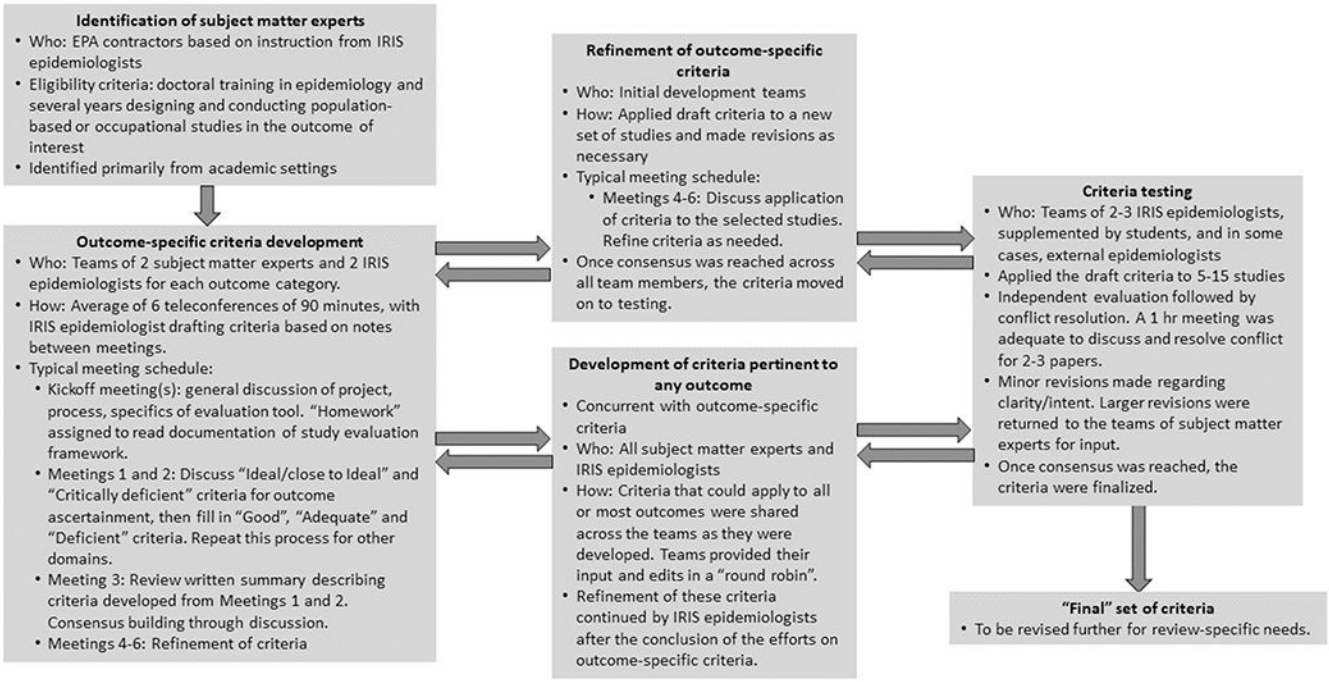


Fig. 2.
Process for outcome-specific criteria development.

Table 1

Questions to guide the development of criteria for each domain in epidemiology studies.

Domain and core question (target bias)	Prompting questions	Follow-up questions
Exposure measurement Does the exposure measure reliably distinguish between levels of exposure in a time window considered most relevant for a causal effect with respect to the development of the outcome? (information bias)	For all: ● Does the exposure measure capture the variability in exposure among the participants, considering intensity, frequency, and duration of exposure? ● Does the exposure measure reflect a relevant time window? If not, can the relationship between measures in this time and the relevant time window be estimated reliably? ● Was the exposure measurement likely to be affected by a knowledge of the outcome? ● Was the exposure measurement likely to be affected by the presence of the outcome (i.e., reverse causality)? For case-control studies of occupational exposures: ● Is exposure based on a comprehensive job history describing tasks, setting, time period, and use of specific materials? For biomarkers of exposure, general population: ● Is a standard assay used? What are the intra- and interassay coefficients of variation? Is the assay likely to be affected by contamination? Are values less than the limit of detection dealt with adequately? ● What exposure time-period is reflected by the biomarker? If the half-life is short, what is the correlation between serial measurements of exposure?	Is the degree of exposure misclassification likely to vary by exposure level? If the correlation between exposure measurements is moderate, is there an adequate statistical approach to ameliorate variability in measurements? If there is a concern about the potential for bias, what is the predicted direction or distortion of the bias on the effect estimate (if there is enough information)?
Outcome ascertainment Does the outcome measure reliably distinguish the presence or absence (or degree of severity) of the outcome? (information bias)	For all: ● Is outcome ascertainment likely to be affected by knowledge of exposure (e.g., consider access to health care, if based on self-reported history of diagnosis)? For case-control studies: ● Is the comparison group without the outcome (e.g., controls in a case-control study) based on objective criteria with little or no likelihood of inclusion of people with the disease? For mortality measures: ● How well does cause of death data reflect occurrence of the disease in an individual? How well do mortality data reflect incidence of the disease? For diagnosis of disease measures: ● Is diagnosis based on standard clinical criteria? If based on self-report of diagnosis, what is the validity of this measure? For laboratory-based measures (e.g., hormone levels): ● Is a standard assay used? Does the assay have an acceptable level of interassay variability? Is the sensitivity of the assay appropriate for the outcome measure in this study population?	Is there a concern that any outcome misclassification is nondifferential, differential, or both? What is the predicted direction or distortion of the bias on the effect estimate (if there is enough information)?
Participant selection Is there evidence that selection into or out of the study (or analysis sample) was jointly related to exposure and to outcome? (selection bias including attrition [primarily], confounding)	For longitudinal cohort: ● Did participants volunteer for the cohort based on knowledge of exposure and/or preclinical disease symptoms? Was entry into the cohort or continuation in the cohort related to exposure and outcome? For occupational cohort: ● Did entry into the cohort begin with the start of the exposure? ● Was follow-up or outcome assessment incomplete, and if so, was follow-up related to both exposure and outcome status? ● Could exposure produce symptoms that would result in a change in work assignment/work status (“healthy worker survivor effect”)? For case-control study: ● Were controls representative of population and time periods from which cases were drawn? ● Are hospital controls selected from a group whose reason for admission is independent of exposure? ● Could recruitment strategies, eligibility criteria, or participation rates result in differential participation relating to both disease and exposure? For population-based survey: ● Was recruitment based on advertisement to people with knowledge of exposure, outcome, and hypothesis?	Were differences in participant enrollment and follow-up evaluated to assess bias? If there is a concern about the potential for bias, what is the predicted direction or distortion of the bias on the effect estimate (if there is enough information)? Were appropriate analyses performed to address changing exposures over time in relation to symptoms? Is there a comparison of participants and nonparticipants to address whether differential selection is likely?

Domain and core question (target bias)	Prompting questions	Follow-up questions
Confounding Is confounding of the effect of the exposure likely? (confounding)	Is confounding adequately addressed by considerations in… a. … participant selection (matching or restriction)? b. … accurate information on potential confounders, and statistical adjustment procedures? c. … lack of association between confounder and outcome, or confounder and exposure in the study? d. … information from other sources? Is the assessment of confounders based on a thoughtful review of published literature, potential relationships (e.g., as can be gained through directed acyclic graphing), minimizing potential overcontrol (e.g., inclusion of a variable on the pathway between exposure and outcome)?	If there is a concern about the potential for bias, what is the predicted direction or distortion of the bias on the effect estimate (if there is enough information)?
Analysis Does the analysis strategy and presentation convey the necessary familiarity with the data and assumptions?[a]	● Are missing outcome, exposure, and covariate data recognized, and if necessary, accounted for in the analysis? ● Does the analysis appropriately consider variable distributions and modeling assumptions? ● Does the analysis appropriately consider subgroups of interest (e.g., based on variability in exposure level or duration, or susceptibility)? ●Is an appropriate analysis used for the study design? ● Is effect modification considered, based on considerations developed a priori? ● Does the study include additional analyses addressing potential biases or limitations (i.e., sensitivity analyses)?	If there is a concern about the potential for bias, what is the predicted direction or distortion of the bias on the effect estimate (if there is enough information)?

[a]The evaluation of the analysis domain is focused on the appropriateness of the analysis rather than any specific form of bias.

Table 2

Key outcome-specific considerations for study evaluation[a].

Category	Outcome	Outcome ascertainment	Exposure (relevant time window)	Population selection	Confounding (variables to consider)	Analysis
Pregnancy outcomes	Preterm birth	● Method used to estimate gestational duration (e.g., early pregnancy dating with ultrasound preferred to birth certificate data)	● Must be measured during pregnancy (preferred) or preconception.	● Timing of study entry (1st trimester preferred).	● Age (u-shaped), gender, pre-pregnancy BMI/adiposity, lifestyle factors (physical activity, diet, smoking, alcohol), SES, co-exposures, pregnancy interval, maternal reproductive and clinical factors	● Outcome treated as dichotomous or three-level variable preferred.
	Spontaneous abortion	● Ascertainment of early loss (through daily urine samples) preferred ● For clinical loss (> 5 weeks gestation), prospective ascertainment of pregnancy and loss preferred to recall or medical records	● Must be prior to outcome, ideally around conception	● Pregnancy planning couples preferred for identifying early loss. ● Population-based or clinic-based with discussion of catchment area preferred to convenience sample ● Recruitment prior to end of first trimester preferred	● Age, smoking, alcohol, parity, gravidity, SES, BMI/adiposity, fertility treatment, pregnancy interval, access to healthcare, co-exposures. If paternal exposure is the focus, maternal covariates should also be considered.	● Outcome treated as dichotomous variable preferred, with survival analysis for prospective designs.
Male reproductive	Reproductive hormones	● Time of collection (morning preferred) ● Information on laboratory assays, quality control procedures	● Can be measured concurrent with outcome.	● Appropriate comparison group (e.g., similar referral patterns)	● Age, SES, smoking, alcohol use, BMI/adiposity, time of day, co-exposures	● Outcome treated as continuous variable preferred.
	Semen parameters (concentration, motility, morphology)	● Semen analysis on one or more samples. ● Manual ascertainment of morphology preferred. ● Abstinence time should be collected and considered. ● Time between collection and analysis (influences motility) ● Information on laboratory methods, quality control procedures	● For adult exposures, exposure measured concurrent with outcome is acceptable. ● Must be before identification of infertility problem unless exposure does not change over time	● Selection at setting other than infertility clinic preferred. ● Sample not limited to volunteers with known fertility problems unless similar selection in comparison group.	● Age, smoking, BMI, chronic disease status, abstinence time, co-exposures	● Outcome treated as continuous or dichotomous (with justified cut-points) variable.
Male or female reproductive	Pubertal development	● Tanner staging or physical measurements (e.g., testicular volume) by trained physician or investigator preferred	● Must be measured before pubertal onset (prenatal or childhood appropriate)	● Population must span the relevant age range (approximately 11–15 years)	● Age, race/ethnicity, SES, diet, obesity, family history of early or late pubertal onset, co-exposures	● Outcome may be treated as time to event (e.g., puberty onset) or ordinal (e.g., Tanner stage)
	Time to pregnancy	● Prospective measurement of couples discontinuing contraception is preferred. ● If recall is used, should be number of cycles or months, and recall time < 10 years is preferred. ● Designs where recall is only among pregnant women or where data is collected in categories (< 3 mo., 3–6 mo., etc.) are Poor.	● Must be prior to conception unless exposure does not vary over time or with pregnancy. ● For women: prior to attempt to conceive, including in utero ● For men: Period	● Couples with no known fecundity impairments strongly preferred. ● Sample not limited to couples that achieved pregnancy. ● Population-based sampling preferred. Other	● Age (maternal and paternal), smoking, BMI, SES, co-exposures	● Outcome treated as time to event preferred.

Category	Outcome	Outcome ascertainment	Exposure (relevant time window)	Population selection	Confounding (variables to consider)	Analysis
		● Assessment of time to pregnancy from medical records (e.g., notation about history of infertility) not acceptable.	of spermatogenesis for attempted conception	sampling should be unbiased (e.g., not infertility clinic).		
Diabetes-related	Diabetes	● Clear definition of diabetes (e.g., ADA); self-report not sole criteria ● Fasting requirements ● Information on laboratory assays, quality control procedures	● Must be measured before diabetes onset.	● Exclude diabetics at baseline ● Cohort followed for sufficient period to allow development of disease	● Age, gender, BMI/ adiposity, lifestyle factors (physical activity, diet, smoking, alcohol), SES, co-exposures	● Outcome treated as dichotomous or three-level variable preferred.
	Insulin resistance	● Absence of diabetes ● HOMA-IR and HOMA-β, or fasting insulin and fasting glucose ● Fasting requirements ● Information on laboratory assays, quality control procedures	● Can be measured concurrent with outcome.	● Exclude individuals with known diabetes	● Age, gender, BMI/ adiposity, lifestyle factors (physical activity, diet, smoking, alcohol), SES, co-exposures	● Outcome treated as continuous variable preferred.
Asthma and allergies	Asthma	● Self-report using validated questionnaires for physician diagnosed asthma incidence or outcome prevalence (e.g., ATS, ISAAC) or medical record review/physician confirmation for incidence; ● Validation of questionnaires in same population preferred.	● Up to two years prior to ascertainment of incidence or concurrent with prevalence or related outcomes	● Population-based sampling preferred; adults or children > 4 years of age. Consider healthy worker effect in occupational studies.	● Age, sex, race, socioeconomic status, season of outcome measurement (for prevalence and related outcomes), geographic location, co-exposures	● Outcome is dichotomous; effect modification by age (child vs adult) preferred if appropriate
	Allergic sensitization and allergy outcomes	● Ascertainment of sensitization via skin prick test or specific immunoglobulin E levels in blood using validated methods; use of standardized cut-points, distinguishing between food and environmental allergens, and testing a panel of 5–10 allergens preferred. ● Ascertainment of site-specific symptoms and outcomes using standardized questionnaires, distinguishing food and other allergy; preferably validated in same population. Ascertainment via self-report of medication use must clarify type of medication.	● Must precede sensitization or symptom onset; or concurrent with prevalence or symptom ascertainment	● Studies among children preferred for sensitization or symptom onset (aged > 4 years). ● Self-reported symptoms preferred over restricting to physician diagnosed allergic disease.	● Age, sex, race, socioeconomic status, season of outcome measurement (for prevalence and related outcomes), geographic location, co-exposures	● Stratified analyses of children and adults; age analyzed as effect modifier for studies of sensitization

Abbreviations: ADA (American Diabetes Association), HOMA-IR (Homeostatic model assessment of insulin resistance), HOMA-β (homeostatic model assessment of β-cell function), ATS (American Thoracic Society), ISAAC (International Society of Arthritis and Allergies in Children).

[a]Classification of the considerations in this table into levels (Good, Adequate, Deficient, and Critically deficient) requires additional considerations that apply to all outcomes and outcome-specific considerations. Rationale for the considerations are discussed in the full outcome-specific evaluation criteria (supplementary materials).

Table 3

Criteria for diabetes and insulin resistance.

Level	Exposure measurement	Outcome ascertainment	Participant selection	Confounding	Analysis
Good 	Will be developed for the specific exposure Diabetes ● Exposure must be measured before diabetes onset (e.g., prospective cohort). Insulin resistance ● Can be measured concurrent with outcome. Criteria that apply to all exposures ● Valid exposure assessment methods used which represent the etiologically relevant time period of interest. ● Exposure misclassification is expected to be minimal.	Diabetes ● For undiagnosed cases, American Diabetes Association (ADA) definition with or without repeated measures, and: o Indicates fasting was required. o Includes information on the laboratory test requirements for hemoglobin A1c and for conducting the oral glucose tolerance tests. o Provides some information on assays, quality control procedures, reliability or validity measures. ● For diagnosed cases, self-reported physician diagnosis or medical treatment for diabetes can be used. Participants without a diagnosis should be tested to avoid false negatives. Insulin resistance ● Homeostatic model assessment (HOMA) of insulin resistance (HOMA-IR) and β-cell function (HOMA-β) in the absence of diabetes from fasting glucose and insulin concentrations measured in plasma. AND/OR ● Fasting insulin and fasting glucose measures – indicates fasting was required. Fasting time accounted for in analysis if compliance was an issue. AND ● Provides some information on assays, quality control procedures, reliability or validity measures.	Diabetes ● Prospective cohort or other design that allows for identification of incident disease. ● Must exclude individuals with diabetes at baseline. Prediabetes can be included but should be addressed in analysis via stratification or sensitivity analysis. ● Cohort followed for sufficient period to allow for development of disease. Insulin resistance ● Cross-sectional studies are appropriate. ● Must exclude individuals with known diabetes. Note for both: for studies of children/adolescents, lack of exclusion of diabetes cases may be acceptable due to low prevalence. Criteria that apply to all outcomes ● Minimal concern for selection bias based on description of recruitment process. ● Exclusion and inclusion criteria specified and would not induce bias. ● Participation rate is reported at all steps of study (e.g., initial enrollment, follow-up, selection into analysis sample). If rate is not high, there is appropriate rationale for why it is unlikely to be related to exposure (e.g., comparison between participants and nonparticipants or other available information indicates differential selection is not likely).	Diabetes and Insulin resistance ● Risk factors that should be considered as possible confounders include age, gender, BMI/adiposity, lifestyle factors (physical activity, diet, smoking, alcohol), SES, co-exposures Criteria that apply to all outcomes Contains all of the following: ● Conveys strategy for identifying key confounders. This may include: a priori biological considerations, published literature, or statistical analyses; with recognition that not all “risk factors” are confounders. ● Inclusion of potential confounders in statistical models not based solely on statistical significance criteria (e.g., $p < 0.05$ from stepwise regression). ● Does not include variables in the models that have been shown to be influential colliders or intermediates on the causal pathway. ● Key confounders are evaluated and considered to be unlikely sources of substantial bias. This will often include: o Presenting the distribution of potential confounders by levels of the exposure and/or the outcomes of interest (with amount of missing data noted); or o Consideration that potential confounders were rare among the study population, or were expected to be poorly correlated with exposure of interest; or o Presenting a progression of model results with adjustment for different potential confounders, including consideration of the function forms of potential confounders, if warranted.	Criteria that apply to all outcomes ● Use of an optimal characterization of the outcome variable. ● Quantitative results presented (e.g., effect estimates and confidence limits). ● Descriptive information about outcome and exposure provided (where applicable): o Amount of missing data noted and addressed appropriately. o For exposure, includes LOD (and percentage less than LOD) and discussion of cut-points and transformations. ● Includes analyses that address robustness of findings (e.g., examination of exposure-response, relevant sensitivity analyses). Effect modification examined based only on a priori rationale with sufficient numbers. ● No deficiencies in analysis evidence. Discussion of some details may be absent (e.g., examination of outliers).

Level	Exposure measurement	Outcome ascertainment	Participant selection	Confounding	Analysis
Adequate 	Will be developed for the specific exposure Diabetes ● Exposure must be measured before diabetes onset. Insulin resistance ● Can be measured concurrent with outcome. Criteria that apply to all exposures ● Valid exposure assessment methods used which represent the etiologically relevant time period of interest. ● Exposure misclassification may exist but is not expected to greatly change the effect estimate.	Diabetes ● ADA definition without repeated measures, and no other information provided. or ● The use of medical records with details on criteria used to define diabetes, in the absence of laboratory tests conducted specifically for the study. Insulin resistance ● Fasting insulin and fasting glucose measures – no details of fasting or laboratory assays provided.	Same as Good, but: Criteria that apply to all outcomes ● Enough of a description of the recruitment process to be comfortable that there is no serious risk of bias. ● Inclusion and exclusion criteria specified and would not induce bias. ● Participation rate is incompletely reported but available information indicates participation is unlikely to be related to exposure.	Criteria that apply to all outcomes Similar to Good, but may not have included all key confounders, or less detail may be available on the evaluation of confounders (e.g., sub-bullets in Good). It is possible that residual confounding could explain part of the observed effect,but concern is minimal.	Criteria that apply to all outcomes Same as Good, except: ● Descriptive information about exposure provided but may be incomplete; might not have discussed missing data, or cut-points, or shape of distribution. or ● Some important analyses that address the robustness of findings are not performed.
Deficient 	Will be developed for the specific exposure Diabetes ● Exposure must be measured before diabetes onset unless the exposure is persistent in the body. Insulin resistance ● Can be measured concurrent with outcome. Criteria that apply to all exposures ● Valid exposure assessment methods used which represent the etiologically relevant time period of interest. There may be concerns about reverse causality, but there is no direct evidence that it is influencing the effect estimate. ● Exposed groups are expected to contain a notable proportion of unexposed or minimally exposed individuals, the method did not capture important temporal or spatial variation, or there is other evidence of exposure misclassification.	Diabetes ● Treatment for diabetes or use of medical records without details on criteria used to define diabetes. or ● Self-reported physician diagnosis is the sole criteria for case ascertainment; participants without a diagnosis are not tested and presumed to be non-cases. Insulin resistance – none defined.	Criteria that apply to all outcomes ● Little information on recruitment process, selection strategy, sampling framework, and/or participation. OR ● Aspects of the recruitment process, selection strategy, sampling framework, or participation raise the potential for bias (e.g., healthy worker effect, survivor bias).	Criteria that apply to all outcomes ● Does not include variables in the models that have been shown to be influential colliders or intermediates on the causal pathway. And any of the following: ● The potential for bias to explain some of the results is high based on an inability to rule out residual confounding, such as a lack of demonstration that key confounders of the exposure-outcome relationship were considered. ● Descriptive information on key confounders (e.g., their relationship relative to the outcomes and exposure levels) are not presented. ● Strategy of evaluating confounding is unclear or is not recommended (e.g., based on statistical significance criteria or stepwise regression only).	Criteria that apply to all outcomes ● Descriptive information about exposure levels not provided (where applicable). or ● Effect estimate presented without standard error or confidence interval. or ● Non-optimal analysis methods used (e.g., correlation instead of linear regression)

Level	Exposure measurement	Outcome ascertainment	Participant selection	Confounding	Analysis
Critically deficient 	Will be developed for the specific exposure Diabetes ● Exposure measure does not reflect exposure before onset of diabetes or is known to be affected by disease status. Criteria that apply to all exposures ● Exposure measurement does not characterize the etiologically relevant period for the outcome or is not valid. ● There is direct evidence that reverse causality could account for the observed association. ● Exposure measurement was not independent of outcome status.	Diabetes ● Studies using self-reported diabetes with no clear information on the questions used to ascertain diabetes status. ● Use of glucosuria to identify diabetes cases. ● Use of diabetes mortality. Insulin resistance ● Reporting HOMA-β in the absence of HOMA-IR Criteria that apply to all outcomes ● Invalid/insensitive marker of outcome. ● Outcome ascertainment was not independent from exposure status.	Criteria that apply to all outcomes ● Aspects of the processes for recruitment, selection strategy, sampling framework, or participation, or specific available data result in concern that bias resulted in a large impact on effect estimates.	Criteria that apply to all outcomes ● Includes variables in the models that have been shown to be influential colliers or intermediates in the causal pathway, indicating that substantial bias is likely from this adjustment. or ● Confounding is likely present and not accounted for, indicating that the results were most likely due to bias (i.e., confounders associated with the outcome and exposure in the study could explain the majority of the reported results).	Criteria that apply to all outcomes ● Results presented as statistically “significant”/ “not significant” or just p-values (i.e., without including effect estimates). or ● Effect modification examined without clear a priori rationale and without providing main effects. or ● Analysis methods were not appropriate for the design or data.