A Section 508–conformant HTML version of this article is available at https://doi.org/10.1289/EHP11532.

Evidence Synthesis of Observational Studies in Environmental Health: Lessons Learned from a Systematic Review on Traffic-Related Air Pollution

Hanna Boogaard,[1] Richard W. Atkinson,[2] Jeffrey R. Brook,[3] Howard H. Chang,[4] Gerard Hoek,[5] Barbara Hoffmann,[6] Sharon K. Sagiv,[7] Evangelia Samoli,[8] Audrey Smargiassi,[9] Adam A. Szpiro,[10] Danielle Vienneau,[11,12] Jennifer Weuve,[13] Frederick W. Lurmann,[14] and Francesco Forastiere[15]

[1]Health Effects Institute, Boston, Massachusetts, USA
[2]Population Health Research Institute, St. George's University of London, London, United Kingdom
[3]Occupational and Environmental Health Division, Dalla Lana School of Public Health, University of Toronto, Toronto, Ontario, Canada
[4]Department of Biostatistics and Bioinformatics, Rollins School of Public Health, Emory University, Atlanta, Georgia, USA
[5]Institute for Risk Assessment Sciences, Environmental Epidemiology, Utrecht University, Utrecht, the Netherlands
[6]Institute for Occupational, Social and Environmental Medicine, Centre for Health and Society, Medical Faculty, University of Düsseldorf, Düsseldorf, Germany
[7]Center for Environmental Research and Children's Health, Division of Epidemiology, University of California Berkeley School of Public Health, Berkeley, California, USA
[8]Department of Hygiene, Epidemiology and Medical Statistics, School of Medicine, National and Kapodistrian University of Athens, Athens, Greece
[9]Department of Environmental and Occupational Health, School of Public Health, University of Montreal, Montreal, Quebec, Canada
[10]Department of Biostatistics, University of Washington, Seattle, Washington, USA
[11]Swiss Tropical and Public Health Institute, Allschwil, Switzerland
[12]University of Basel, Basel, Switzerland
[13]Department of Epidemiology, Boston University School of Public Health, Boston, Massachusetts, USA
[14]Sonoma Technology, Inc., Petaluma, California, USA
[15]Environmental Research Group, School of Public Health, Imperial College London, London, United Kingdom

Background: There is a long tradition in environmental health of using frameworks for evidence synthesis, such as those of the U.S. Environmental Protection Agency for its Integrated Science Assessments and the International Agency for Research on Cancer Monographs. The framework, Grading of Recommendations Assessment, Development, and Evaluation (GRADE), was developed for evidence synthesis in clinical medicine. The U.S. Office of Health Assessment and Translation (OHAT) elaborated an approach for evidence synthesis in environmental health building on GRADE.

Methods: We applied a modified OHAT approach and a broader "narrative" assessment to assess the level of confidence in a large systematic review on traffic-related air pollution and health outcomes.

Discussion: We discuss several challenges with the OHAT approach and its implementation and suggest improvements for synthesizing evidence from observational studies in environmental health. We consider the determination of confidence using a formal rating scheme of up- and downgrading of certain factors, the treatment of every factor as equally important, and the lower initial confidence rating of observational studies to be fundamental issues in the OHAT approach. We argue that some observational studies can offer high-confidence evidence in environmental health. We note that heterogeneity in magnitude of effect estimates should generally not weaken the confidence in the evidence, and consistency of associations across study designs, populations, and exposure assessment methods may strengthen confidence in the evidence. We mention that publication bias should be explored beyond statistical methods and is likely limited when large and collaborative studies comprise most of the evidence and when accrued over several decades. We propose to identify possible key biases, their most likely direction, and their potential impacts on the results. We think that the OHAT approach and other GRADE-type frameworks require substantial modification to align better with features of environmental health questions and the studies that address them. We emphasize that a broader, "narrative" evidence assessment based on the systematic review may complement a formal GRADE-type evaluation. https://doi.org/10.1289/EHP11532

Introduction

Evidence synthesis is widely used to summarize findings of health effects studies of environmental exposures. Such syntheses are typically part of systematic reviews of observational epidemiologic study findings and often include meta-analyses, with adapted methods first developed by the Cochrane Collaboration for use in clinical medicine.[1] Evidence integration brings together multiple data streams (e.g., observational epidemiologic studies, human and animal experiments, and *in vitro* studies),[2,3] thereby expanding the overall evaluation of the strength of the evidence for causality determination and risk assessment.

Frameworks for evidence synthesis and integration confer structure, consistency, and transparency to these processes. There is a long tradition in environmental health of using such frameworks. Those currently in use are based on frameworks proposed in the 1960s, including the 1964 report of the U.S. Surgeon General on Smoking and Health[4] and the 1965 landmark paper by Sir Austin Bradford Hill, The Environment and Disease: Association or Causation?[5] Five of the nine viewpoints proposed by Hill (strength of association, consistency, specificity, temporality, and coherence) were also employed in the Surgeon General report.[4]

The Monographs program of the International Agency for Research on Cancer (IARC)[3,6] has played an important role in applying Hill's viewpoints for cancer risk assessment as well as the U.S. Environmental Protection Agency (EPA) cancer guidelines.[7,8] Since its inception in 1971, IARC periodically updated its general procedures for scientific review and evaluation for cancer hazard identification. Recent milestones include an update and elaboration of its systematic review process, such as more attention to the quality and informativeness of epidemiological studies, including their exposure assessment methods,[3,6]

Address correspondence to Hanna Boogaard. Email: jboogaard@healtheffects.org
Supplemental Material is available online (https://doi.org/10.1289/EHP11532).
F. W. L. is employed by Sonoma Technology Inc. All panel members served as consultants, paid by the Health Effects Institute, on the research herein.
Received 9 May 2022; Revised 12 September 2023; Accepted 25 October 2023; Published 22 November 2023.

and the transparent management of potential conflicts of interests.[9,10]

In the United States, the EPA Office of Research and Development has developed its Integrated Risk Information System (IRIS) program since 1985 to guide human health assessments for hazard identification and risk assessment of chemicals in the environment. The methods of the IRIS program continue to evolve, with a major decision to begin implementing systematic review methods following National Research Council recommendations.[11,12] The IRIS office articulated its systematic review methods in a handbook, which was recently reviewed by the National Academies.[13,14] Another EPA office, the Office of Pollution Prevention and Toxics, has developed its own systematic review approaches under the requirements of the revised Toxic Substances Control Act (TSCA).[15] Its methods were recently reviewed by a separate Committee of the National Academies, independent of the IRIS evaluation.[16] Both IRIS and TSCA use formal schemes to rate risk of bias in individual studies and aggregate the ratings within or across different evidence streams.

Additionally, the Office of Research and Development at EPA developed a weight-of-evidence approach that it uses in its Integrated Science Assessments (ISA) to determine causality, which inform decisions on the National Ambient Air Quality Standards. This weight-of-evidence approach, first applied in 2008 and evolving since[17] is described in detail in all recent ISAs and elsewhere.[2] The framework supports the consistent, transparent evaluation of evidence across multiple evidence streams and determination of causality. While it considers all of the relevant aspects of evidence synthesis and integration, the framework is flexible; it does not quantitatively rate study quality and does not use formal rating approaches for causal determination. Recently, another Committee of the National Academies evaluated this weight-of-evidence approach, concluding that the "fundamental structure of the weight-of-evidence approach described in the 2015 ISA Preamble[17] allows effective determination of causality for both health and welfare effects." The Committee recommended making the process and determinations more transparent—but not the use of formal rating schemes.[18]

In 2000, the Grading of Recommendations Assessment, Development, and Evaluation (GRADE) working group was established to construct a framework for developing clinical guidelines.[19,20] In applying this framework, guideline developers initially group available studies by key study design features, rate the groupings, and then up- or downgrade them based on specific factors, such as risk of bias. The framework offers a unified approach for comparing the effectiveness of an intended beneficial treatment in the clinical realm with that of a standard treatment and for evaluating evidence to develop clinical guidelines for therapeutic interventions. GRADE has been widely applied. It was adopted by the Word Health Organization (WHO) for guideline development in 2012,[21] irrespective of whether guidelines apply to questions of therapeutic interventions in clinical medicine, preventive interventions in public health, or to assess potential harms from environmental exposures.

In contrast to clinical research, environmental health research investigates the health effects of a potentially harmful environmental exposure, one often experienced over years or decades. The GRADE environmental health working group adapted the original GRADE approach to environmental health to guide on assessing the "certainty" of evidence on the health effects of environmental exposures.[22] The "certainty" of evidence (or "quality" or "confidence") in GRADE reflects the extent of our confidence that the effect estimates are correct or the certainty that a true effect lies on one side of a specified threshold or within a chosen range.[23] An adapted GRADE framework was recently applied to evaluate the epidemiologic evidence in the WHO guidelines for environmental noise and air quality.[24,25]

Other approaches, including the Navigation Guide,[26] and the framework from the Office of Health Assessment and Translation (OHAT)[27] were developed to specifically address environmental health questions. Both frameworks are based on Cochrane Collaboration and GRADE but with distinct modifications, such as the integration of human and animal studies. Both frameworks are also applied widely. For example, WHO and the International Labor Organization adopted the Navigation Guide to evaluate occupational burden of disease since 2016.[28,29]

All frameworks described above require expert judgement and confer structure, consistency, and transparency to those judgements. It has long been known that a more transparent and systematic approach is superior to expert judgement alone in synthesizing evidence.[30]

In a large systematic review of traffic-related air pollution (TRAP) and health outcomes, we systematically assessed the epidemiologic evidence. For this, we applied a modified OHAT approach to assess the level of confidence in the quality of the body of evidence. In addition, we applied a broader, "narrative" assessment to determine the level of confidence in the presence of an association. The resulting Health Effects Institute (HEI) Special Report, a short communication paper, and three papers on TRAP and selected health outcomes (mortality, stroke, and diabetes) were recently published.[31–35] Here, we describe lessons learned in synthesizing evidence from observational studies of TRAP. Though the paper builds on the HEI Special Report,[31] new insights, clarifications, and discussions have been included in response to peer review. We expect this reflection to be useful for future assessments of environmental exposures and health.

Methods

General Methods of the Systematic Review on TRAP and Health

Following its well-cited 2010 critical review,[36] HEI appointed a new expert panel to systematically evaluate the epidemiological evidence regarding the associations between long-term exposure (months to years) to TRAP and selected adverse health outcomes. The panel consisted of 13 experts in epidemiology, exposure assessment, and statistics at institutions in North America and Europe.

The panel used a systematic approach to search the literature, select studies for inclusion in the review, assess study quality, summarize results, and reach conclusions about the confidence in the association between TRAP and a specific health outcome. To this end, a review protocol was published in 2019[37] and registered in Prospero.[38]

Health outcomes were selected by the panel based on evidence of causality (causal or likely causal) according to the latest determination for general air pollution (broader than TRAP) from available authoritative integrated science assessments,[39–42] and other considerations such as relevance for public health and policy, and resources available. The panel selected clinical outcomes (rather than preclinical and biomarker measures), including birth outcomes (e.g., term low birth weight and preterm birth), respiratory outcomes (e.g., asthma onset), cardiometabolic outcomes (e.g., ischemic heart disease and diabetes), and all-cause and cause-specific (e.g., circulatory and respiratory) mortality.[31]

A PECOS (Population, Exposure, Comparator, Outcome, and Study) statement was developed, along with inclusion and exclusion criteria for each PECOS domain in relation to the selected health effects of long-term exposure to TRAP. The focus of the review was on health effects observed in the general population. Cohort, case-control, cross-sectional, and intervention studies using individual-level health outcome data were eligible.[31]

A new exposure framework was developed to determine whether a study was sufficiently specific to TRAP. The panel included studies that evaluated exposure to nitrogen dioxide (NO_2), elemental carbon (EC), ultrafine particles (UFP), $PM_{2.5}$ and PM_{10} (particles smaller than 2.5 and 10 μm, respectively), and other pollutants and indirect traffic measures (distance to major roads and traffic density). Studies were subject to additional (e.g., spatial resolution) criteria developed to ensure the exposure was sufficiently specific to TRAP.[31]

An extensive search was conducted of literature published between January 1980 and July 2019. Two reviewers checked studies for inclusion eligibility according to the PECOS statement. Data from all included studies were extracted and evaluated extensively. To represent the associations specific to the TRAP mixture, effect estimates from single-pollutant models (as opposed to multipollutant models) were selected for the meta-analysis. Random-effects meta-analysis was used when at least three estimates were available for a specific exposure-outcome pair. Risk of bias was assessed for all exposure-outcome associations that were included in the meta-analyses, using a modified version of the risk of bias tool developed for the systematic reviews informing the WHO Air Quality Guidelines.[25,43] In brief, the risk of bias tool guides the assessment of each study across six domains: *a*) confounding, *b*) selection bias, *c*) exposure assessment, *d*) outcome measurement, *e*) missing data, and *f*) selective reporting. Most domains have subdomains. A rating for each subdomain and an overall rating per domain were derived using three categories (low/moderate/high). No summary classification was derived across the domains.[43]

Where possible, the panel performed additional meta-analyses to assess the consistency of the association, for example, across geographic regions, within time periods, by level of risk of bias per domain, and with more extensive adjustment for individual-level smoking.[31]

Evidence Synthesis

The panel was charged with transparently assessing the level of confidence in associations between TRAP and selected adverse health outcomes. The panel assessed confidence in *a*) the quality of the body of evidence using a modified OHAT approach and *b*) the presence of an association between TRAP and the selected adverse health outcomes using a "narrative" assessment. The rationale and the methods are reported below.

The panel did not assess causality because it did not conduct separate, independent systematic assessments of the mechanistic, toxicological, and human clinical studies relating TRAP to human health.

Modified OHAT assessment. We chose the OHAT approach as a guide for the assessment of the confidence in the quality of the body of evidence.[27,44] While the OHAT approach was developed based on the original GRADE, we considered the OHAT approach better suited than the GRADE approach that was applied to evaluate the epidemiologic evidence in the WHO guidelines for environmental noise and air quality.[24,25] For instance, OHAT provides a more robust rationale for the initial confidence of observational studies (see below) and includes an upgrade factor for consistency of the results across populations and study design.

In short, available studies on a particular health outcome are initially grouped by key study design features (i.e., controlled exposure, exposure prior to the outcome, individual outcome data, and a comparison group). Each grouping of studies is given an initial confidence rating (high, moderate, low, very low) by those study design features (not study-by-study). This initial confidence rating for the body of evidence from each group of studies is then downgraded for factors that decrease confidence in the body of evidence (risk of bias, unexplained inconsistency, indirectness, imprecision, and publication bias) and upgraded for factors that increase confidence in the body of evidence (large magnitude of effect, exposure-response, consistency, and consideration of residual confounding or other factors that increase the confidence in the body of evidence).[27]

The OHAT approach directly translates the three highest confidence ratings in the quality of the body of evidence (high, moderate, or low) into the level of evidence in support of the presence of an adverse health effect. The confidence rating, very low, translates to a level-of-evidence conclusion of inadequate evidence.[27] This direct translation is problematic: specifically, the panel determined that, before drawing conclusions about an effect based on a confidence rating, additional relevant factors should be scrutinized, such as the number and size of the studies, direction and magnitude of the association, the consistency of the results from the meta-analyses and the studies not meta-analyzed, and the generalizability of the findings. There is an additional consideration that precludes direct translation, namely the situation where no adverse health effects are detected. Hence, in OHAT, only the conclusion "evidence of no effect" is reached when there is a high confidence in the body of evidence.[27] Conceptually, the panel thought that it is particularly problematic to evaluate an evidence base to exclude a potential environmental health risk and to support a conclusion of "evidence of no effect" because several additional features of the studies should be considered, such as sufficient time elapsed between exposure and the outcome, adequate exposure contrast, and time-windows for exposure and outcome.[45] These considerations are not fully captured in the OHAT approach according to the panel. Hence, the panel restricted the formal OHAT confidence assessment to a rating of the quality of the body of evidence and added a separate complementary and broader "narrative" assessment of the confidence in the presence of an association (see section "Narrative assessment").

The panel also slightly modified the OHAT approach. We refer to Supplemental Material, "List of main modifications to OHAT for the traffic review" for the main modifications. In contrast to OHAT guidance,[27] the panel gave all types of cohort studies (not only prospective) and case-control studies based on incident cases an initial rating of moderate because three key study design features were often present (exposure precedes the outcome, individual-level data, and a comparison group). Similar to the OHAT approach,[27] the panel started with an initial rating of low confidence for cross-sectional studies because one cannot typically assert that the exposure precedes the outcome. Ecologic studies were excluded from consideration in the traffic review. Note that in original GRADE guidance,[20] all observational studies start at low confidence, but this disregards typical and potentially critical differences in quality across observational study designs.

We did not use two grading factors—indirectness and large magnitude of effect—in the process of downgrading and upgrading of confidence in the body of evidence. Indirectness was not applicable because we included only studies of human exposure to TRAP in direct association with the health outcomes. A large magnitude of effect was unlikely to be meaningful based on experiences in the systematic reviews informing the WHO Air Quality Guidelines, where large or very large effect sizes [i.e., large relative risk (RR) >2 or very large RR >5 as defined in OHAT] never occurred.[46,47] Large RRs were not observed in our review either. We note that the use of large RRs was first proposed as part of "good epidemiological practices" guidelines in the sound science movement advocated by industry in the 1990s.[48] Though used as an upgrading factor in OHAT (instead of attempting to discredit epidemiologic studies with lower RRs[48]) we contend that this factor is problematic and

not very useful for most environmental epidemiology studies, where relative risks are often small.

As TRAP is a complex mixture, the panel evaluated the body of evidence separately for each exposure indicator (e.g., NO_2 and EC) included in the review. The panel then evaluated the body of evidence across all included traffic-related air pollutants to obtain an assessment of the confidence in the quality of the body of evidence for TRAP. In this assessment, the panel also used evidence from studies of indirect traffic measures, such as distance to major roadways and traffic density and other results that did not enter a meta-analysis, such as those involving categorized exposures or involving traffic-related air pollutants with fewer than three studies. For example, very few studies were identified for some pollutants, in particular nontailpipe PM indicators and UFP.

"Narrative" assessment. The panel developed its "narrative" assessment in response to limitations and challenges experienced when a related GRADE approach was applied in the systematic reviews informing the WHO Air Quality guidelines[46,47] and difficulties when developing the modified OHAT protocol (described above). The "narrative" assessment accompanied and complemented the modified OHAT assessment in evaluating the level of confidence in the presence of an association of TRAP with adverse health effects. As anticipated in the review protocol, the panel elaborated on some methods when the review was already underway. All expansions, including the "narrative" assessment, were based solely on methodological considerations and were independent of study results.

The "narrative" confidence assessment was also performed on the body of evidence (not study-by-study) and based on the systematic review. Results of the meta-analysis (summary estimates, forest plot, heterogeneity of the estimates) supplemented with studies not included in the meta-analysis formed the foundation of this assessment. We then evaluated many of the same factors related to the internal validity of the studies addressed in the OHAT approach, but we inspected additional factors too (Table S1). We elaborate on some of those sometimes-subtle differences in the "Lessons learned" section. In short, the "narrative" approach gave a prominent weight to the following factors: evaluation of the number of studies, their (variability in) location, and sample size; risk of bias of the individual studies, including traffic noise for some outcomes; the magnitude and direction of the association; a monotonic exposure-response function; and consistency of study findings across populations, age groups, time periods, study designs, and pollutants and the generalizability of study results. For example, associations that were replicated in several studies across different populations, across several pollutants, or that used different epidemiological approaches were more likely to represent a true association than isolated observations from small, single studies.

In summary, in evaluating the level of confidence that TRAP is associated with the selected health outcome, the broader, "narrative" assessment considered all evidence in the systematic review without using a formal rating scheme.

Overall confidence. Both approaches (modified OHAT and "narrative" assessment) yielded confidence ratings of high, moderate, low, or very low. Subsequently, we combined the findings from both assessments into an overall confidence assessment. In case of agreement, the overall assessment was the same as the individual assessments (e.g., two assessments of "high" resulted in "high" overall); if not in agreement we have indicated both (e.g., "moderate to high"), since the panel considered both assessments complementary, reflecting the complex issues in determining the level of confidence.

For elaboration on the methods for evidence synthesis, we refer to HEI's Special Report[31] including "Additional Materials 5.3."

Main Findings of the Systematic Review on TRAP and Health

The panel found a high level of confidence that associations exist between long-term exposure to TRAP and early death from all causes and circulatory diseases. A moderate-to-high level of confidence was found for associations of TRAP with lung cancer mortality, asthma onset in children and adults, and acute lower respiratory infections in children. The panel's confidence in the evidence for an association was considered moderate for respiratory mortality, asthma ever and active asthma in children, term low birth weight, small for gestational age, ischemic heart disease events, and diabetes morbidity. The confidence in all other selected health outcomes was low-to-moderate (e.g., stroke morbidity and mortality), low (e.g., preterm birth), or very low-to-low (e.g., acute lower respiratory infections in adults). In total, 353 studies were included in the systematic review, with dozens of exposure indicators and outcomes. We refer to HEI's Special Report for all of the findings.[31]

Discussion

Lessons Learned

The experience of conducting the traffic review provided insights on the processes that are discussed below and informed our suggested improvements to key elements of the OHAT approach and, more widely, to processes for synthesizing evidence from observational studies in environmental health. Table 1 contains a summary.

Evidence Synthesis Needs a Broader, "Narrative" Approach to Maximize What Can Be Learned from Observational Studies in Environmental Health

Although a key strength of the OHAT approach is that it makes judgements transparent, the panel noted several challenges and limitations with OHAT methods and its implementation. Among the challenges were the formal rating scheme of up- and downgrading of certain factors, the equal weight of all factors, some of the tools to decide upon up- or downgrades, and that because of those tools, the evaluation can be heavily geared toward studies entering a meta-analysis.

The formal OHAT approach indicates up- and downgrading factors to be considered. The panel found these factors to be logical but had issues with the system of adding and subtracting the score per each factor to the initial judgment (e.g., the initial moderate plus one upgrade minus two downgrades results in a final confidence rating of low). A judgment per each factor requires an expert evaluation, and therefore, it is critical that sufficient text should accompany each judgement to fully appreciate the assessment. While OHAT does not preclude such text in its handbook,[27] its importance could be emphasized.

A second challenge is that the OHAT approach treats all factors as equally important. This is problematic, e.g., we argue that risk of bias may be a more severe issue than publication bias in relatively large evidence bases of exposures studied for an extended period, such as air pollution and noise. We acknowledge that there is no straightforward solution for this weighting issue. This debate is similar to the risk of bias discussion, where we refrained from developing a summary classification across the different domains of risk of bias.

Hence, in determining the level of confidence, the panel deemed it necessary to accompany the modified OHAT assessment with a broader, "narrative" approach to fully capture some of the important and complex nuances that would have otherwise been missed by conducting either alone. Such a "narrative" approach is likely to be useful for future assessments, even as GRADE-type approaches are evolving to better align with features of environmental health

Table 1. Suggested improvements to key elements of the OHAT approach[27] for synthesizing evidence from observational studies in environmental health.

Key elements of the OHAT approach	Suggested improvements by the Traffic Review Panel[a]
Evidence synthesis	
Use a GRADE-type approach to assess confidence in the quality of the body of evidence.	Complement the GRADE-type assessment with a broader, "narrative" approach to maximize what can be learned from observational studies in environmental health.
Assign an initial low or moderate level of confidence to all types of observational studies.	Consider that in environmental health, where randomized controlled trials are generally not appropriate, some observational studies can offer high-confidence evidence.
Assess the statistical heterogeneity of results and downgrade the confidence rating if substantial heterogeneity is found.	Sources of heterogeneity can strengthen or weaken the confidence in the evidence and should be carefully explored. Some heterogeneity is expected in studies of the health effects of environmental exposures, due to different populations, locations, and study settings. Consider primarily the direction of the effect estimate rather than its magnitude. Because different methods and study designs that generate similar findings may strengthen the confidence, a separate upgrading factor for consistency should be added to GRADE.[20]
Assess publication bias using Egger's test and funnel plots and downgrade accordingly.	Publication bias is not necessarily expected when large and collaborative (multicenter) studies comprise most of the evidence and/or if evidence has accrued over several decades. Use additional approaches to explore the possibility of publication bias.
Risk of bias in individual studies	
Compare study with randomized controlled trials or hypothetical target experiment as ideal study.	Do not consider randomized controlled trials as ideal study.
Evaluate bias in different domains (e.g., confounding, selection bias, measurement error).	Focus on identifying the most likely influential sources of bias—based on methodologic and subject matter expertise—classifying each study on the basis of how effectively it has addressed each potential bias and determining whether results differ across studies in relation to each hypothesized source of bias.
Rate potential biases (e.g., low, moderate, high) using a risk of bias tool.	Rate biases considering the suggestions in the row above. Those ratings should not be used to dismiss studies based on bias but to conduct sensitivity analyses comparing findings from studies of high bias and low/moderate bias.

Note: GRADE, Grading of Recommendations Assessment, Development, and Evaluation; OHAT, U.S. Office of Health Assessment and Translation.
[a]HEI appointed an expert Panel to systematically evaluate the epidemiological evidence regarding the associations between long-term exposure (months to years) to TRAP and selected adverse health outcomes. The Panel consisted of 13 experts in epidemiology, exposure assessment, and statistics at institutions in North America and Europe, and are coauthors of this paper.

questions and the studies that address them. We acknowledge that our "narrative" approach could have been elaborated more for increased transparency. Examples of more detailed narrative approaches are the preambles of IARC and ISA of the EPA.[6,17]

We further note that the factors considered in the "narrative" approach broadly overlap with those included in the OHAT approach but differ in how they are considered and weighted. As an example, if multiple studies are conducted in very diverse populations, heterogeneity in effect estimates is likely to be high. This could conceivably result in a decrease of confidence using the OHAT approach for unexplained inconsistency, depending on whether the consideration "diverse populations" is considered sufficient explanation of the heterogeneity. Using the "narrative" approach, we would argue that if associations are predominantly positive in those diverse populations, this would increase confidence.

For the most part, in our review, the OHAT and "narrative" assessments reached the same confidence conclusions regarding the body of evidence and the presence of an association between TRAP and the selected health outcomes, although there were some exceptions (Table S2 and S3). Discrepancies between the two assessments emerged typically when studies that did not enter a meta-analysis provided additional information on highly traffic-specific pollutants. We describe some examples below. Figure S1 and Table S4 contain a complete assessment using both approaches for asthma onset in children, and we refer readers to HEI's Special Report for all other assessments.[31]

Asthma outcomes. In children, the panel examined evidence on the association between TRAP and asthma onset (incidence), asthma ever (lifetime prevalence), and active asthma (last 12 months prevalence). For NO_2, the most studied pollutant, the modified OHAT approach indicated high confidence in the body of evidence for asthma onset (with an initial moderate rating for cohort studies that was upgraded for a monotonic exposure-response) and moderate confidence for asthma ever and active asthma (with an initial low rating for cross-sectional studies that was upgraded for consistency across populations in both outcomes). No downgrades were applied for NO_2. The high or moderate OHAT rating for NO_2 (n ranged from 12 to 21 studies), resulted in similar ratings for TRAP, even though the confidence assessment was low and very low for the other less-studied traffic pollutants, including NO_x (n ranged from 3 to 6) and EC (n ranged from 3 to 5).[31]

The "narrative" assessment underscored that a sizable number of well-designed large cohort studies (of asthma onset) and cross-sectional studies (of asthma ever and active asthma) were set in a variety of locations. There were at least two large studies showing a monotonic exposure-response function for NO_2 and asthma incidence. Associations were also reported for some studies not entering a meta-analysis. The summary estimate for NO_2 was positive for all three outcomes, although only borderline significant for asthma onset. Furthermore, summary estimates were positive for most other pollutants, but many were notably imprecise, and all were based on far fewer studies. Given the imprecision in the summary estimates, the panel found that uncertainties remained regarding the association between TRAP and the three asthma outcomes in children and concluded in the "narrative" assessment that there was moderate evidence. Thus, the OHAT and "narrative" assessment reached the same confidence conclusions for asthma ever and active asthma, but the OHAT assessment was one level higher for asthma onset. In other words, imprecision of the effect estimates was an important factor in the "narrative" assessment.[31]

In adults, the OHAT assessment was one level lower than the "narrative" assessment for TRAP and asthma onset. The modified OHAT assessment was moderate, with no up- or downgrades applied, and primarily based on studies of NO_2 ($n = 7$), as too few studies were available for meta-analyses of other traffic pollutants.

The summary estimate for NO_2 and asthma onset in adults was positive with confidence intervals (CIs) that excluded unity. The "narrative" assessment underlined several studies documenting a positive association and also fully incorporated another large and well-designed cohort study that used NO_2 categories. Although this study's analysis of NO_2 categories (rather than a continuous exposure) excluded it from meta-analysis, its results revealed positive associations across progressively higher NO_2 categories. Positive NO_2 associations were reported across different populations, and the few studies on other traffic pollutants like NO_x, PM_{10} and $PM_{2.5}$, and EC also reported positive associations. Hence, in its "narrative" assessment, the panel concluded that the evidence of an association between exposure to TRAP and asthma onset in adults was high.[31]

Stroke incidence. Similarly for stroke incidence, the OHAT assessment was one level lower than the "narrative" assessment. The panel's rating using the modified OHAT method was low. This assessment was based on summary effect estimates that were positive for EC ($n=6$), PM_{10} ($n=5$), and $PM_{2.5}$ ($n=4$) and null for NO_2 ($n=7$) and NO_x ($n=8$). The initial confidence rating was moderate for all five meta-analyzed pollutants due to the cohort and case-control study designs. But all pollutants' ratings (except that of NO_x) were downgraded for imprecision since the CIs of the meta-analytic estimates were wide despite large overall sample sizes and clearly included unity. PM_{10} and $PM_{2.5}$ received upgrades because at least two large studies documented a monotonic exposure-response function. The final ratings using OHAT were low for NO_2 and EC and moderate for NO_x, PM_{10}, and $PM_{2.5}$. Note that there was a moderate confidence for NO_x, which showed null associations. When judging the consistency across pollutants, the confidence was downgraded to low because of inconsistencies across the pollutants and because $PM_{2.5}$ and PM_{10} studies were only moderately specific to traffic.[31]

In contrast, the panel found a moderate level of confidence based on the "narrative" assessment because several studies that did not enter a meta-analysis provided additional evidence in support of a positive association, including two well-designed studies that were highly specific to traffic. Two studies yielded positive associations with indirect traffic measures. Substantial confounding by noise was ruled out in four studies with available noise exposure data. Moreover, there was some concern regarding potential bias toward the null because of overadjustment and inclusion of covariates in a few large studies with negative or null associations, which carried substantial weight in the meta-analyses.[31]

All-cause mortality. For all-cause mortality, both assessments yielded high confidence ratings, which were based on slightly different although complementary reasoning. The modified OHAT approach was based on summary effect estimates indicating positive associations for NO_2 ($n=11$), NO_x ($n=5$), EC ($n=11$), PM_{10} ($n=6$), $PM_{2.5}$ ($n=12$), copper (Cu) ($n=3$), and iron (Fe) ($n=3$). The initial confidence rating was moderate due to the cohort study design. The panel derived high-confidence judgements for NO_2, EC, and $PM_{2.5}$; moderate for NO_x and PM_{10}; and low for Cu and Fe. These judgements stemmed from a combination of downgrades for risk of bias (Cu) and imprecision (NO_x, PM_{10}, and Fe) and upgrades for a monotonic exposure-response (all pollutant except Cu and Fe) and consistency across populations (NO_2).[31]

The "narrative" assessment also yielded a high rating and entailed a sizable number of well-conducted cohort studies, several of them in very large study populations. Studies were conducted in a larger number of locations by different research groups. In addition to the consistently positive associations found in the meta-analyses, associations in individual studies were consistent in direction but differed in magnitude, as expected. Further support came from studies on indirect traffic measures and from studies excluded from meta-analysis, notably, studies using highly traffic-specific exposure estimates. Regarding the internal validity of the studies, the panel noted that most of the results were adjusted for major potential confounders. Moreover, the associations were positive in different locations, indicating an unlikely influence of confounding on the body of evidence, as the associations between TRAP exposure and lifestyle/socioeconomic factors have been shown to differ in direction, depending on study area. Hence, insufficient adjustment for confounders may have resulted in both upward and downward bias across studies. Further increasing confidence were studies reporting TRAP associations that remained after adjusting for traffic noise.[31]

Observational Studies Can Offer High-Confidence Evidence in Environmental Health

GRADE[20] considers randomized controlled trials as the gold standard for evaluating health effects and assigns observational studies a low initial confidence. This logic is derived from evaluation of clinical research questions that focus on the intended beneficial effect of a medical treatment. In a similar vein, human controlled trials and experimental animal studies in the OHAT approach[27] typically receive a high initial confidence rating because they meet the key study design feature, controlled exposure, whereas all of the other study designs receive a lower rating.

Randomized controlled trials or other experimental studies are often not appropriate or ethical for studies of potentially harmful health effects of environmental exposures (intervention studies being the exception). Furthermore, randomized controlled trials often involve short follow-up times and limited sample sizes. In contrast, investigations of environmental exposure effects may require follow-up over many years to capture long etiologic induction periods and necessitate very large sample sizes (up to millions) due to relatively small effect sizes, which, although less relevant for clinical decision-making, are quite relevant to public health decision-making.[49] Finally, randomized controlled trials often have limited generalizability, because they recruit highly selected samples of persons meeting stringent criteria—healthier and with fewer underlying conditions—than the population that might eventually use the treatment.[49] By contrast, a large epidemiological study in the general population can include the full spread of people at-risk (e.g., people with preexisting diseases, people in different life stages) of health effects in response to environmental exposures.

These fundamental differences motivate rethinking the approach for evaluating the confidence in a body of evidence in environmental health. Since randomized controlled trials are largely inappropriate for validly answering environmental health questions, observational studies that are designed to minimize systematic error can rise to the occasion. Therefore, for future assessments, the panel recommends that observational studies, specifically cohort and case-control studies of incident cases, start with a high confidence rating. This recommendation does not extend to cross-sectional and ecologic studies. In GRADE-type assessments, the confidence rating for a body of evidence from observational studies could be downgraded if substantial biases are likely that would affect the effect estimates significantly. The panel prefers the approach of explicitly describing biases in a body of evidence over automatically assigning lower initial confidence to all observational studies.

Consider All Relevant Studies in Evidence Synthesis

The inclusion of relevant studies should be comprehensive, and all studies should be judged based on their scientific merit. A systematic review may involve the conduct of meta-analyses; however, studies included in a meta-analysis often represent a subset

of the available studies. In the traffic review,[31] only about half of all studies considered were included in meta-analyses. For example, the panel did not pursue meta-analyses of indirect traffic measures, such as distance to major roadways and traffic density and studies on traffic-specific PM fractions, because the varying definitions across the studies precluded such quantitative analyses. Yet, those studies contained information on the potential health effect of near-road traffic and were thus used in the confidence assessment.

The panel emphasizes that meta-analyses do not automatically increase confidence in the evidence, and studies not fitting into a statistical summary may be equally informative and merit inclusion in evidence synthesis. Apparent consistency in the results of a meta-analysis and the individual studies results contributing to it may reflect consistent biases. A meta-analytic estimate might incorporate a small subset of highly informative studies that are overwhelmed by a large number of weaker ones. Meta-analyses can even obscure informative heterogeneity based on varying methods.[50]

The OHAT approach[27] does not fundamentally limit the assessment to studies included in a meta-analysis. However, when we applied the tools to decide upon up- or downgrades, we noted that in studies not included in a meta-analysis, some factors were difficult to evaluate, specifically imprecision and unexplained inconsistency, both of which relied on quantitative judgements. Hence, because of those tools, the evaluation could still be heavily geared toward studies entering a meta-analysis.

The panel decided to not apply the OHAT approach to exposure-outcome pairs that lacked meta-analyses. Within the OHAT approach, those non-meta-analyzed studies were considered when the panel evaluated the body of evidence across all included traffic-related air pollutants and indirect traffic measures to obtain a confidence assessment for TRAP. Results from studies that did not enter meta-analyses were mainly considered in the broader, "narrative" approach. Although the limitations of focusing on meta-analyzable studies have been recognized in GRADE,[51] and some solutions have been proposed in the absence of a summary estimate,[52,53] more work is clearly needed to avoid overemphasis of meta-analyses results in evidence synthesis.

Heterogeneity of the Magnitude in Effect Estimates among Observational Studies Should Generally Not Be Used to Downgrade Confidence

The Panel noted that sources of heterogeneity can strengthen or weaken confidence in the evidence and should be carefully explored. It is critical to distinguish between heterogeneity arising from true differences in associations (effect modification) from heterogeneity arising from methodological differences across studies. In the absence of a systematic error, the magnitude of the effect estimate could vary substantially across different populations and locations, with different exposures assessed, pollution mixtures or co-pollutants, time periods, age structure, and follow-up times, for example. Thus, some heterogeneity is expected in estimated health effects of environmental exposures. Furthermore, experts generally have higher confidence in an association if studies have found associations in multiple, diverse locations and populations, factors all likely leading to (true) heterogeneity.

In the traffic review,[31] the panel *a priori* identified subgroups of interest for sensitivity analyses, provided there were sufficient studies to conduct meta-analyses. Subgroups were defined by geographical area, time period, high vs. lower risk of bias per domain of the risk of bias tool, and confounder adjustment for individual-level smoking.

No single statistical measure of consistency of findings across studies is ideal, and statistical tests for heterogeneity have well-known limitations.[54] Moreover, they are less reliable when the number of studies is small. In the traffic review, downgrading because of unexplained inconsistency was considered if heterogeneity was high (operationalized as $I^2 > 75\%$) and applied after reviewing the potential sources of heterogeneity, including risk of bias, and considering the direction of the effect estimate rather than its magnitude. Of note, inconsistency was less of a concern for a group of studies all reporting positive associations, albeit with inconsistent magnitude, as the purpose of the assessment was to identify the presence of an association rather than to estimate its magnitude. This purpose may differ for other applications in environmental health. The OHAT approach[27] is not entirely consistent on this issue, as it lists both magnitude and direction as possible reasons for downgrades but also offers an example (in Table 11 of its handbook), suggesting that high and significant heterogeneity is not a serious concern if all studies have effect estimates in the same direction. We very much support the latter interpretation.

The OHAT approach[27] uses consistency as an extra upgrading factor and so have almost all frameworks for review and evaluation of environmental hazards and risks to inform policy. GRADE[20] should also consider adding such an upgrading factor. Traditionally, the consistency of associations across study designs, populations, and exposure assessment methods provides additional confidence in the results. Recently, the usefulness of consistency has gained support in the concept of triangulation, albeit for a different purpose.[55] The underlying premise is that if different epidemiological approaches, possibly with unrelated sources of bias, all support the same conclusion, the confidence in the evidence is strengthened. This is particularly compelling when the key sources of bias of some of the approaches are predicted to influence estimates in opposite directions.[56] For these reasons, the panel upgraded the confidence when results were based on different study designs (cohort studies/case-control vs. cross-sectional studies) that supported the same conclusions. Likewise, the panel also upgraded for consistency of associations across large geographic areas, as the potential for bias was judged to be different in different populations. The decision to upgrade was not always obvious, and the panel did not upgrade, for example, in the case of only a few studies or with studies consistently reporting null findings.

Publication Bias Should Be Explored beyond Statistical Methods

The OHAT aproach[27] suggests that some degree of publication bias is likely, and downgrading should be reserved for instances where the concern is serious enough to significantly reduce confidence in the body of evidence. We used funnel plots and Egger's regression tests to help assess publication bias, provided there were at least 10 studies in the meta-analysis. However, even 10 studies may be too few, because the results of the Egger tests also depend on the study size and magnitude of associations.[57,58] Most importantly, true heterogeneity in effect size unrelated to publication bias may also lead to asymmetrical funnel plots and statistically significant Egger tests.

In the traffic review,[31] the Egger test was highly significant for EC and total mortality but not for NO_2 and $PM_{2.5}$. One small study of EC reported a relatively large effect estimate, which also had the widest CI; furthermore, studies with relatively wide CIs reported both significant and nonsignificant findings. The panel judged that the observed asymmetry in EC was more likely due to heterogeneity than to publication bias and did not downgrade for publication bias. Moreover, the panel noted that 7 of the 11 EC studies also reported a NO_2 estimate for which the Egger test was nonsignificant. It is difficult to imagine a scenario in which the publication bias mechanism is stronger for EC than for NO_2

and $PM_{2.5}$ studies. The panel *a priori* did not expect that publication bias would be a major issue in the group of cohort studies, given the effort required to perform (multicenter) cohort studies, often including collaboration between different research groups including cohort owners, environmental epidemiologists, statisticians, and exposure scientists, an argument also made in a recent systematic review underpinning the WHO Air Quality Guidelines.[46] The air pollution body of evidence has accrued over several decades and includes several large studies. Thus, publication bias in such a body of evidence is likely limited compared with a body of evidence of relatively recent studies with early small positive studies.[27]

Hence, the panel noted that statistical methods for publication bias should be applied with caution, and it is important to assess consistency of the tests across the body of evidence (e.g., comparing across pollutants). Other approaches may be useful. Examples include a subgroup analysis of multicenter studies with single city studies or an analysis of differences in effect estimates from earlier vs. later studies. The latter was also explored by the panel to detect early positive studies of a small sample size. From this analysis, there was no clear sign of publication bias, and overall, this downgrading factor was never applied in the traffic review.

Assessing the Influence of Specific Sources of Potential Bias instead of Using a Risk of Bias Tool

A critical step in the systematic review process is assessing the risk of bias in included studies. Although various tools exist, there is no consensus about the best approach for assessing risk of bias in observational studies.[59–61] The panel used an adapted version of the risk of bias tool developed for the systematic reviews informing the WHO Air Quality Guidelines,[43] because the tool was designed for assessment of risk of bias in observational air pollution studies.

The panel compared effect estimates in subgroups of studies rated as high vs. moderate or low risk of bias for specific domains, such as confounding or selection bias. When effect estimates from studies with low or moderate risk of bias were virtually the same as those with high risk of bias, the panel did not downgrade the evidence. In this setting, all studies were included in the overall assessment. When effect estimates from studies with low or moderate risk of bias were considerably different from effect estimates of studies at high risk of bias (whatever the direction of the difference) and there were sufficient studies in the low or moderate categories, we omitted the high risk of bias studies from the confidence assessment. Downgrading occurred only when effect estimates from studies at low/moderate risk of bias were considerably different from estimates from studies at high risk of bias and the body of the evidence of studies with low or moderate risk of bias was limited. This could apply to few studies and/or a small weight in the meta-analysis of all studies. Those steps concord with recent guidance on integrating the risk of bias assessments of individual studies into evidence synthesis in the context of observational studies in the OHAT approach and GRADE.[22,27]

After comparing findings between subgroups within a risk of bias domain, the panel applied the downgrading for risk of bias in 18% of the meta-analyses (in total, 87 meta-analyses were conducted for the selected health outcomes). Most of the downgrades were applied for birth outcomes: 12 of the 19 meta-analyses on birth outcomes were downgraded based on risk of bias. Only four downgrades were made for the other outcomes. High risk of bias studies were never excluded from the confidence-rating phase entirely.

In the traffic review,[31] most of the studies in meta-analyses were rated as low to moderate risk of bias for all but the "confounder" domain. For this domain, about one-third of the studies were rated as high risk of bias because important confounders were not adjusted for. Differences in effect estimates between the low/moderate and high risk of bias studies were small, and hence, no downgrade was applied. To evaluate risk of bias for confounding, the panel developed a list of important confounders, which had to be adjusted for to be judged as low risk of bias. The list was based on subject matter-informed directed acyclic graphs (DAGs) and included age, sex, individual level or neighborhood socioeconomic status (SES), body mass index (BMI), and smoking. BMI was not included for respiratory mortality and morbidity outcomes, similar to the WHO risk of bias guidance,[43] and sex was not included for birth outcomes. The definition of an important confounder is to some degree subjective because confounding by a specific factor may differ widely between study populations and settings. For example, the association between TRAP exposure and potential confounders, such as lifestyle/socioeconomic factors have been shown to differ in direction, depending on study area.[31] Moreover, risk of bias indicates the potential for the results of an individual study to be biased and does not inform on actual bias in a particular study. If in a specific study area, there is no association between exposure and a specific covariate, failure to adjust for this covariate does not result in bias. Neither does a score of moderate or high risk of bias inform about the size of a potential bias; for example, while risk of bias can be high due to a methodologic problem, the size of the actual bias might be very small and vice versa. A high risk of bias determination also does not indicate the direction of bias, which can vary according to specific study conditions, with the potential for different biases to operate in countervailing directions.

The panel cautions about applying strict evaluation criteria, "formulaic" approaches, checklists, comparisons to a hypothetical target experiment, and the creation of an overall study quality rating in risk of bias assessments. The panel also cautions about excluding studies based on risk of bias. Instead, the panel recommends using the results of subgroup analysis by risk of bias per domain. Eick et al.[60] compared different risk of bias tools in a case study relevant for environmental health and also cautioned about the use of an overall study quality rating and exclusion based on risk of bias. Furthermore, the panel agreed with Eick et al.[60] that completing the detailed risk of bias evaluations of each study is time-consuming. Any approach, however, to assess risk of bias is likely time-consuming to do justice to the complexity of epidemiological studies.

For future assessments, the panel advocates that bias assessments should focus more on identifying the most likely influential sources of bias for all relevant studies—based on methodologic and subject matter expertise—classifying each study on the basis of how effectively it has addressed each potential bias and determining whether results differ across studies in relation to each hypothesized source of bias, as described in Savitz et al.[61] Such an approach can provide insight into the potential influence of each specific bias, identify a subset of studies likely to best approximate the true association, and suggest features needed to improve future research. The approach fits into the concept of triangulation, as discussed in the section "Heterogeneity of the magnitude in effect estimates among observational studies should generally not be used to downgrade confidence."

Imprecision Needs a Better Definition

Guidance from the OHAT approach[27] on imprecision is based on the 95% CI of either individual study estimates or the meta-analytical summary estimates and on the optimal information size (OIS) criterion. Furthermore, the OHAT approach defines that estimates are generally considered imprecise for ratio measures (e.g., odds ratio) when the ratio of the upper to lower 95%

CI for most studies is ≥10. The panel found this too lenient and nondiscriminant for the air pollution database. We used a much stricter definition: A narrow (precise) CI was defined as a difference on the log scale of ≤0.1 from the upper to lower 95% CI. A wide (imprecise) CI was defined as a difference on a log scale of >0.1 between the upper and lower 95% CI. We did not downgrade in case of a 95% CI not including unity, as in this case, the imprecision is not sufficiently large to affect the overall interpretation. The latter is in line with GRADE guidance.[20] Finally, the 95% CI of a meta-analytical summary estimate from a random effects model is affected by individual study precision and heterogeneity. We observed that the 95% CI of some individual studies was smaller than that of the summary estimate. This consideration is important because we should not downgrade twice for imprecision and unexplained inconsistency.

Downgrading or Upgrading Should Be Treated Independently

Another important choice in the application of the OHAT approach was whether upgrades in confidence should be assigned without consideration to downgrades that have been assigned and vice versa. The panel opted to evaluate the downgrading and the upgrading factors independently, following the GRADE application in the systematic reviews informing the WHO guidelines for environmental noise and air quality.[24,25,46,47] There may be some clear exceptions; for example, if a downgrade for risk of bias has been made, one should not upgrade for large magnitude of the effect. This specific lesson learned was not applicable to the "narrative" assessment because no formal rating system was used.

Conclusions

Based on a large evidence synthesis of epidemiological studies of TRAP, we have described several challenges with the OHAT approach and its implementation. We have suggested improvements to the OHAT approach and, more broadly, to processes for synthesizing evidence from observational studies in environmental health.

We think that the OHAT approach and other GRADE-type frameworks require substantial modification to align better with features of environmental health questions and the studies that address them. We note that more applications of different environmental stressors and across multiple evidence streams (e.g., observational epidemiologic studies, human and animal experiments, and *in vitro* studies) are needed to further develop those frameworks. We emphasize that a broader, "narrative" evidence assessment based on the systematic review may complement a formal GRADE-type evaluation and may maximize what can be learned from observational studies in environmental health.

Acknowledgments

HEI is indebted to the panel, the consultants to the panel, external peer reviewers, and contract team for their expertise, cooperation, and enthusiasm. We specifically thank J. Fussell, F. Kelly, T. Nawrot, and G. Wellenius as consultants to the panel, and B. Brunekreef for chairing the peer review process. We also would like to thank M. Kutlar Joss and R. Kappeler for literature searches and data extraction. In addition, we would like to thank A. da Silveira Fleck, P. Haddad, L. Hoffmann, L. Stucki, M. Sadoine, Z. Roth, and E. Wüthrich for their help with data extraction. We would like to thank the following HEI Science Staff for their contribution: A. Patton, D. Crouse, E. van Vliet, M. Ondras, E. Tanner, D. Greenbaum, R. O'Keefe, R. Shaikh, and A van Erp.

Research described in this article was conducted under contract to the HEI, an organization jointly funded by the United States Environmental Protection Agency (assistance award no. CR-83998101) and certain motor vehicle and engine manufacturers. The views expressed in this article are those of the authors and do not necessarily reflect the views of the Health Effects Institute or its sponsors.

References

1. Chalmers I, Dickersin K, Chalmers TC. 1992. Getting to grips with Archie Cochrane's agenda. BMJ 305(6857):786–788, PMID: 1422354, https://doi.org/10.1136/bmj.305.6857.786.
2. Owens EO, Patel MM, Kirrane E, Long TC, Brown J, Cote I, et al. 2017. Framework for assessing causality of air pollution-related health effects for reviews of the national ambient air quality standards. Regul Toxicol Pharmacol 88:332–337, PMID: 28526659, https://doi.org/10.1016/j.yrtph.2017.05.014.
3. Samet JM, Chiu WA, Cogliano V, Jinot J, Kriebel D, Lunn RM, et al. 2020. The IARC monographs: updated procedures for modern and transparent evidence synthesis in cancer hazard identification. J Natl Cancer Inst 112(1):30–37, PMID: 31498409, https://doi.org/10.1093/jnci/djz169.
4. U.S. HEW (United States Department of Health, Education, and Welfare). 1994. *Smoking and Health: Report of the Advisory Committee to the Surgeon General of the Public Health Service*. Washington, DC: U.S. Department of Health, Education, and Welfare, Public Health Service, Center for Disease Control.
5. Hill AB. 1965. The environment and disease: association or causation? Proc R Soc Med 58(5):295–300, PMID: 14283879.
6. IARC (International Agency for Research on Cancer). 2019. *IARC Monographs on the Evaluation of Carcinogenic Hazards to Humans. Preamble*. Lyon, France: International Agency for Research on Cancer.
7. U.S. EPA (U.S. Environmental Protection Agency). 1986. Guidelines for carcinogen risk assessment. Federal Register 51(185):33992–34003.
8. U.S. EPA (U.S. Environmental Protection Agency). 2005. Guidelines for Carcinogen Risk Assessment. EPA/630/P-03/001F. Washington, DC: U.S. Environmental Protection Agency.
9. Cogliano V, Baan R, Straif K, Grosse Y, Secretan B, El Ghissassi F, et al. 2005. Transparency in IARC monographs. Lancet Oncol 6(10):747, https://doi.org/10.1016/S1470-2045(05)70380-6.
10. IARC (International Agency for Research on Cancer). 2006. *IARC Monographs on the Evaluation of Carcinogenic Risks to Humans. Preamble*. Lyon, France: International Agency for Research on Cancer.
11. NRC (National Research Council). 2011. *Review of the Environmental Protection Agency's Draft Integrated Risk Information System (IRIS) Assessment of Formaldehyde*. Washington, DC: The National Academies Press.
12. NRC (National Research Council). 2014. *Review of Environmental Protection Agency's Integrated Risk Information System (IRIS) Process*. Washington, DC: The National Academies Press.
13. NASEM (National Academies of Sciences, Engineering, and Medicine). 2022. *A Review of U.S. EPA's ORD Staff Handbook for Developing IRIS Assessments: 2020 Version*. Washington, DC: The National Academies Press.
14. U.S. EPA (U.S. Environmental Protection Agency). 2020. *ORD Staff Handbook for Developing IRIS Assessments (Public Comment Draft, Nov 2020)*. Washington, DC: U.S. EPA Office of Research and Development.
15. U.S. EPA (U.S. Environmental Protection Agency). 2018. *Application of Systematic Review in TSCA Risk Evaluations*. Washington, DC: Office of Chemical Safety and Pollution Prevention.
16. NASEM (National Academies of Sciences, Engineering, and Medicine). 2021. *The Use of Systematic Review in EPA's Toxic Substances Control Act Risk Evaluations*. Washington, DC: The National Academies Press.
17. U.S. EPA (U.S. Environmental Protection Agency). 2015. *Preamble to the Integrated Science Assessments. EPA/600/R-15/067*. Research Triangle Park, NC: U.S. EPA.
18. NASEM (National Academies of Sciences, Engineering, and Medicine). 2022. *Advancing the Framework for Assessing Causality of Health and Welfare Effects to Inform National Ambient Air Quality Standard Reviews*. Washington, DC: The National Academies Press.
19. Guyatt GH, Oxman AD, Vist GE, Kunz R, Falck-Ytter Y, Alonso-Coello P, et al. 2008. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. BMJ 336(7650):924–926, PMID: 18436948, https://doi.org/10.1136/bmj.39489.470347.AD.
20. Schünemann H, Brożek J, Guyatt G, Oxman AD, editors. 2013. *GRADE Handbook for Grading Quality of Evidence and Strength of Recommendations*. Hamilton, ON: The GRADE Working Group.
21. WHO (World Health Organization). 2014. *WHO Handbook for Guideline Development*, 2nd ed. Geneva, Switzerland: WHO.
22. Morgan RL, Thayer KA, Bero L, Bruce N, Falck-Ytter Y, Ghersi D, et al. 2016. GRADE: assessing the quality of evidence in environmental and occupational

health. Environ Int 92–93:611–616, PMID: 26827182, https://doi.org/10.1016/j.envint.2016.01.004.
23. Hultcrantz M, Rind D, Akl EA, Treweek S, Mustafa RA, Iorio A, et al. 2017. The GRADE working group clarifies the construct of certainty of evidence. J Clin Epidemiol 87:4–13, PMID: 28529184, https://doi.org/10.1016/j.jclinepi.2017.05.006.
24. WHO (World Health Organization). 2018. *Environmental Noise Guidelines for the European Region*. Geneva, Switzerland: World Health Organization Regional Office for Europe.
25. WHO (World Health Organization). 2021. *WHO Global Air Quality Guidelines: Particulate Matter (PM2.5 And PM10), Ozone, Nitrogen Dioxide, Sulfur Dioxide and Carbon Monoxide*. Geneva, Switzerland: World Health Organization.
26. 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. Environ Health Perspect 122(10):1007–1014, PMID: 24968373, https://doi.org/10.1289/ehp.1307175.
27. OHAT (Office of Health Assessment and Translation). 2019. *Handbook for Conducting a Literature-Based Health Assessment Using OHAT Approach for Systematic Review and Evidence Integration*. Washington, DC: National Toxicology Program, National Institute of Environmental Health Sciences, U.S. Department of Health and Human Services.
28. Pega F, Momen NC, Ujita Y, Driscoll T, Whaley P. 2021. Systematic reviews and meta-analyses for the WHO/ILO joint estimates of the work-related burden of disease and injury. Environ Int 155:106605, PMID: 34051644, https://doi.org/10.1016/j.envint.2021.106605.
29. WHO/ILO (World Health Organization and International Labour Organization). 2021. *WHO/ILO Joint Estimates of the Work-Related Burden of Disease and Injury, 2000–2016: Technical Report with Data Sources and Methods*. Geneva, Switzerland: World Health Organization and the International Labour Organization.
30. Antman EM, Lau J, Kupelnick B, Mosteller F, Chalmers TC. 1992. A comparison of results of meta-analyses of randomized control trials and recommendations of clinical experts. Treatments for Myocardial Infarction. JAMA 268(2):240–248, PMID: 1535110.
31. HEI (Health Effects Institute). 2022. *Systematic Review and Meta-Analysis of Selected Health Effects of Long-Term Exposure to Traffic-Related Air Pollution. Special Report 23*. Boston, MA: HEI.
32. Boogaard H, Patton AP, Atkinson RW, Brook JR, Chang HH, Crouse DL, et al. 2022. Long-term exposure to traffic-related air pollution and selected health outcomes: a systematic review and meta-analysis. Environ Int 164:107262, PMID: 35569389, https://doi.org/10.1016/j.envint.2022.107262.
33. Boogaard H, Samoli E, Patton AP, Atkinson RW, Brook JR, Chang HH, et al. 2023. Long-term exposure to traffic-related air pollution and non-accidental mortality: a systematic review and meta-analysis. Environ Int 176:107916, PMID: 37210806, https://doi.org/10.1016/j.envint.2023.107916.
34. Haddad P, Kutlar Joss M, Weuve J, Vienneau D, Atkinson RW, Brook JR, et al. 2023. Long-term exposure to traffic-related air pollution and stroke: a systematic review and meta-analysis. Int J Hyg Environ Health 247:114079, PMID: 36446272, https://doi.org/10.1016/j.ijheh.2022.114079.
35. Kutlar Joss M, Boogaard H, Samoli E, Patton AP, Atkinson RW, Brook JR, et al. 2023. Long-term exposure to traffic-related air pollution and diabetes: a systematic review and meta-analysis. Int J Public Health 68:1605718, PMID: 37325174, https://doi.org/10.3389/ijph.2023.1605718.
36. HEI (Health Effects Institute). 2010. *Traffic-Related Air Pollution: A Critical Review of the Literature on Emissions, Exposure, and Health Effects. Special Report 17*. Boston, MA: HEI.
37. HEI (Health Effects Institute). 2019. *Protocol for a Systematic Review and Meta-Analysis of Selected Health Effects of Long-Term Exposure to Traffic-Related Air Pollution*. Boston, MA: HEI. https://www.healtheffects.org/system/files/TrafficReviewProtocol.pdf [accessed 7 November 2023].
38. Boogaard H, Patton A, Forastiere F, Lurmann F, Atkinson R, Brook J, et al. 2019. *Systematic Review and Meta-Analysis of Selected Health Effects of Long-Term Exposure to Traffic-Related Air Pollution*. https://www.crd.york.ac.uk/prospero/display_record.php?ID=CRD42019150642 [accessed 7 November 2023].
39. Health Canada. 2016. *Human Health Risk Assessment for Diesel Exhaust*. Ottawa, ON: Health Canada.
40. IARC (International Agency for Research on Cancer). 2016. *Outdoor Air Pollution: IARC Monographs on the Evaluation of Carcinogenic Risks to Humans*, vol.109. Lyon, France: IARC.
41. U.S. EPA (U.S. Environmental Protection Agency). 2016. *Integrated Science Assessment for Oxides of Nitrogen–Health Criteria. EPA/600/R-15/068*. Washington, DC: U.S. EPA.
42. U.S. EPA (U.S. Environmental Protection Agency). 2019. *Integrated Science Assessment for Particulate Matter (Final Report December 2019). EPA/600/R-19/188*. Washington, DC: U.S. EPA.
43. WHO (World Health Organization). 2020. *Risk of Bias Assessment Instrument for Systematic Reviews Informing WHO Global Air Quality Guidelines*. Geneva, Switzerland: WHO.
44. Rooney AA, Boyles AL, Wolfe MS, Bucher JR, Thayer KA. 2014. Systematic review and evidence integration for literature-based environmental health science assessments. Environ Health Perspect 122(7):711–718, PMID: 24755067, https://doi.org/10.1289/ehp.1307972.
45. Cooper GS, Lunn RM, Ågerstrand M, Glenn BS, Kraft AD, Luke AM, et al. 2016. Study sensitivity: evaluating the ability to detect effects in systematic reviews of chemical exposures. Environ Int 92–93:605–610, PMID: 27156196, https://doi.org/10.1016/j.envint.2016.03.017.
46. Chen J, Hoek G. 2020. Long-term exposure to PM and all-cause and cause-specific mortality: a systematic review and meta-analysis. Environ Int 143:105974, PMID: 32703584, https://doi.org/10.1016/j.envint.2020.105974.
47. Huangfu P, Atkinson R. 2020. Long-term exposure to NO_2 and O_3 and all-cause and respiratory mortality: a systematic review and meta-analysis. Environ Int 144:105998, PMID: 33032072, https://doi.org/10.1016/j.envint.2020.105998.
48. Ong EK, Glantz SA. 2001. Constructing "sound science" and "good epidemiology": tobacco, lawyers, and public relations firms. Am J Public Health 91(11):1749–1757, PMID: 11684593, https://doi.org/10.2105/ajph.91.11.1749.
49. Steenland K, Schubauer-Berigan MK, Vermeulen R, Lunn RM, Straif K, Zahm S, et al. 2020. Risk of bias assessments and evidence syntheses for observational epidemiologic studies of environmental and occupational exposures: strengths and limitations. Environ Health Perspect 128(9):95002, PMID: 32924579, https://doi.org/10.1289/EHP6980.
50. Savitz DA, Forastiere F. 2021. Do pooled estimates from meta-analyses of observational epidemiology studies contribute to causal inference? Occup Environ Med 78(9):621–622, PMID: 34158356, https://doi.org/10.1136/oemed-2021-107702.
51. Norris SL, Bero L. 2016. GRADE methods for guideline development: time to evolve? Ann Intern Med 165(11):810–811, PMID: 27654340, https://doi.org/10.7326/M16-1254.
52. Murad MH, Mustafa RA, Schünemann HJ, Sultan S, Santesso N. 2017. Rating the certainty in evidence in the absence of a single estimate of effect. Evid Based Med 22(3):85–87, PMID: 28320705, https://doi.org/10.1136/ebmed-2017-110668.
53. Thayer KA, Schünemann HJ. 2016. Using GRADE to respond to health questions with different levels of urgency. Environ Int 92–93:585–589, PMID: 27126781, https://doi.org/10.1016/j.envint.2016.03.027.
54. Rücker G, Schwarzer G, Carpenter JR, Schumacher M. 2008. Undue reliance on I^2 in assessing heterogeneity may mislead. BMC Med Res Methodol 8:79, PMID: 19036172, https://doi.org/10.1186/1471-2288-8-79.
55. Lawlor DA, Tilling K, Davey Smith G. 2016. Triangulation in aetiological epidemiology. Int J Epidemiol 45(6):1866–1886, PMID: 28108528, https://doi.org/10.1093/ije/dyw314.
56. Pearce N, Vandenbroucke JP, Lawlor DA. 2019. Causal inference in environmental epidemiology: old and new approaches. Epidemiology 30(3):311–316, PMID: 30789434, https://doi.org/10.1097/EDE.0000000000000987.
57. Lin L, Chu H. 2018. Quantifying publication bias in meta-analysis. Biometrics 74(3):785–794, PMID: 29141096, https://doi.org/10.1111/biom.12817.
58. Murad MH, Chu H, Lin L, Wang Z. 2018. The effect of publication bias magnitude and direction on the certainty in evidence. BMJ Evid Based Med 23(3):84–86, PMID: 29650725, https://doi.org/10.1136/bmjebm-2018-110891.
59. Bero L, Chartres N, Diong J, Fabbri A, Ghersi D, Lam J, et al. 2018. The risk of bias in observational studies of exposures (ROBINS-E) tool: concerns arising from application to observational studies of exposures. Syst Rev 7(1):242, PMID: 30577874, https://doi.org/10.1186/s13643-018-0915-2.
60. Eick SM, Goin DE, Chartres N, Lam J, Woodruff TJ. 2020. Assessing risk of bias in human environmental epidemiology studies using three tools: different conclusions from different tools. Syst Rev 9(1):249, PMID: 33121530, https://doi.org/10.1186/s13643-020-01490-8.
61. Savitz DA, Wellenius GA, Trikalinos TA. 2019. The problem with mechanistic risk of bias assessments in evidence synthesis of observational studies and a practical alternative: assessing the impact of specific sources of potential bias. Am J Epidemiol 188(9):1581–1585, PMID: 31145434, https://doi.org/10.1093/aje/kwz131.