ADDITIONAL MATERIALS AVAILABLE ON THE HEI WEBSITE

Special Report 23

Systematic Review and Meta-analysis of Selected Health Effects of Long-Term Exposure to Traffic-Related Air Pollution

HEI Panel on the Health Effects of Long-Term Exposure to Traffic-Related Air Pollution

Chapter 5: General Methods

These Additional Materials were not formatted or edited by HEI. This document was part of the HEI Panel's review process.

Correspondence concerning the Special Report may be addressed to Dr. Hanna Boogaard at Health Effects Institute, 75 Federal Street, Suite 1400, Boston, Massachusetts, 02110; email: *jboogaard@healtheffects.org*.

Although this document was produced with partial funding by the United States Environmental Protection Agency under Assistance Award CR–83467701 to the Health Effects Institute, it has not been subjected to the Agency's peer and administrative review and therefore may not necessarily reflect the views of the Agency, and no official endorsement by it should be inferred. The contents of this document also have not been reviewed by private party institutions, including those that support the Health Effects Institute; therefore, it may not reflect the views or policies of these parties, and no endorsement by them should be inferred.

Chapter 5: General Methods

Additional Materials

5.1 Quality control of screening and data extraction

5.1.*1 Overview*

Extensive quality checks beyond what was anticipated and outlined in the protocol were performed during reference screening and data extraction (Chapter 5). These checks included duplicate screening of references, discussion of disagreements over whether a specific study should be included and checks of the final set of studies and extracted data. In addition, a reliability study of duplicate data extracted was conducted on a subset of the studies (see results below).

5.1.2 Screening

Automated filters were used to exclude studies of short-term or time-series studies or occupational exposure, studies of traffic accidents or those studying protective devices, controlled trials and case crossover studies, and studies in mice and rats (Appendix 5B Search Strategy). 2402 studies out of 13,660 were excluded by filters, and 10% of those studies were randomly selected and manually checked to confirm that they had not been excluded in error.

The initial screening of 10,775 titles and abstracts in Distiller was performed in duplicate by two screeners following a stepwise approach in sets of around 200 to 500 references. After each set, all disagreements were discussed within the contractor team to come to agreement on inclusion or exclusion before continuing. The initial kappa value of agreement was 0.89 for the first set of references and improved over time. In cases where it was not clear based on the title and abstract whether a study met the inclusion criteria, the contractor team included it for full text screening.

All reasons for exclusion at the full-text review stage were reported (Additional Materials 7.3). The main reason for exclusion of each study excluded at this stage was checked by two HEI staff members and discussed with Panel members as necessary.

5.1.3 Data extraction

The contractor team conducted full double entry for about 70 studies (including all mortality studies) to ensure high quality. Comparison of doubly input data showed generally identical results, although it did reveal some ambiguities in some fields in the data extraction form. For example, data extractors had different interpretations of how to consistently extract results for categorical effect estimates, when to extract results from sensitivity analysis, and how to indicate that a confounder (e.g., smoking) was considered in sensitivity analysis.

Inconsistencies identified by initial duplicate data extraction were addressed by further developing and clarifying the data extraction manual and instructing data extractors to bring any questions to discussion within the contractor team. Minimal data extraction of basic study results by members of the contractor team was complemented by information provided by the more senior members of the contractor team to ensure data extraction quality. In addition, the contractors consulted HEI staff and the Panel when expert input was required, for example when defining respiratory outcomes and selecting data to extract when multiple models were presented in the same paper. Finally, Panel members and HEI staff checked the final data in all summary tables and figures during the writing of the report chapters.

As a further check on data extraction accuracy, data extractors entered effect estimates and confidence intervals into DistillerSR in duplicate for the subset of 55 studies that included all-cause and cause-specific mortality outcomes. The vast majority of extracted effect estimates and confidence intervals were identical between two data extractors; of the 175 data points extracted in duplicate, there were 9 disagreements on data extraction from 4 studies (Figure 1). Reasons for disagreements were digit transposition or similar errors in transcription (2 studies) and disagreement about interpreting the labeling of results or which

estimates should have been extracted to best meet the inclusion criteria (2 studies). Disagreements were corrected, and all data were checked multiple times during the preparation of chapters. The coding of pollutants and outcomes was also checked by HEI staff and Panel members for each effect estimate included in summary tables.

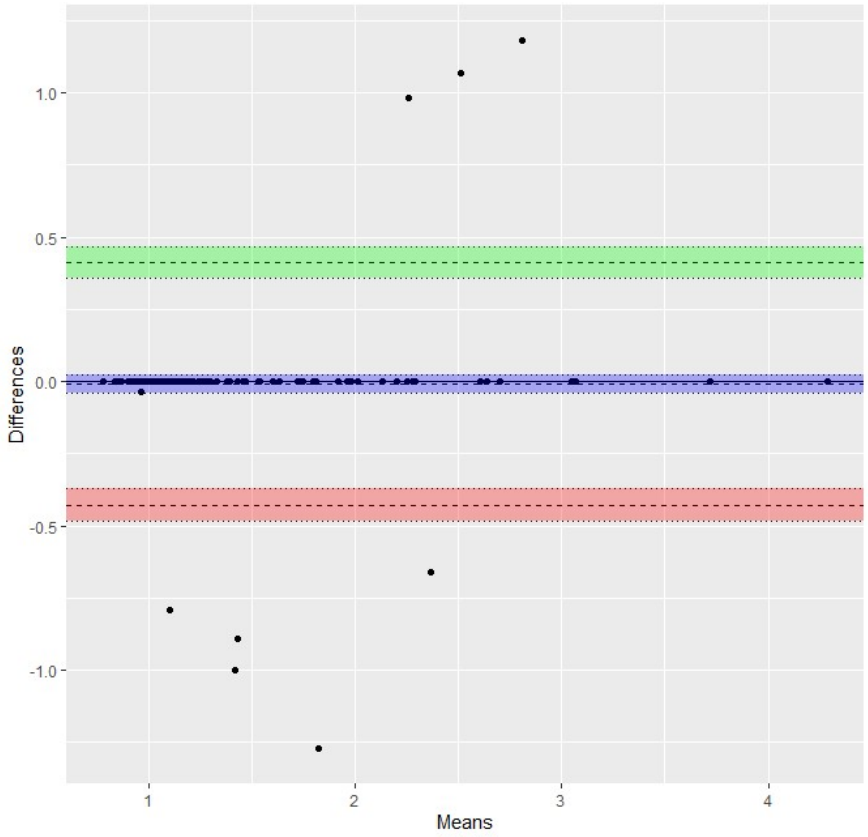

Figure 1. Bland-Altman plot for comparison of effect estimates extracted by two data extractors. Mean bias between data extractors was -0.0081 (95% CI: -0.0400, 0.0238), the lower limit of agreement was -0.427 (95% CI: -0.482, -0.372), and the upper limit of agreement was 0.401 (95% CI: 0.356, 0.466).

5.2 Modified risk of bias tool*.

Subdomain	Low risk criteria (Ideal Study)	Moderate risk criteria	High risk Criteria	Summary of guidance from Traffic Panel to aid interpretation
1.1 Were all confounders considered adjusted for in the design or analysis?	All potential important confounders adjusted for.	Not all potential important confounders adjusted for but with support (e.g., exploratory analysis) of minimal risk due to residual confounding (i.e., there is evidence that this confounder might not lead to severe confounding).	Less than all potential important confounders adjusted for without support (e.g., exploratory analysis) of minimal risk due to residual confounding.	Low: at least all potential important confounders included either in the analysis or in the design: - Age - Sex (not important for birth outcomes) - iSES or nSES - Smoking - BMI or related measures of obesity/physical activity (not important for respiratory mortality, respiratory outcomes) Moderate - Not all included but additional support High - Not all included, no additional support
1.2 Validity of measuring of confounding factors	All potential important confounders measured with documented valid methods.	Not all potential important confounders were measured with documented valid methods; however, there is evidence that this does not lead to severe confounding.	Any potential important confounder not validly assessed AND evidence of residual confounding.	Low: - self-reported age, sex, iSES, smoking, BMI - Administrative information for nSES Moderate - BMI and smoking in medical records or other administrative data - nSES by self-report High - Self-report of smoking and BMI after outcome has occurred
1.3 Control in analysis	Authors used appropriate analysis methods or study designs that controlled for confounding domains.	Authors used inappropriate methods or designs when adjusting for potential important confounders; however, there is evidence that this does not lead to severe confounding.	Authors used inappropriate methods or designs when adjusting for potential important confounders.	Note: this question is independent from Question 1.1 and 1.2. Low: all other cases and using the following methods: - Randomization - Restriction by design or subgroup analysis - Direct adjustment - Standardization - Matching - Use of propensity score - Stratification and Mantel-Haenszel-estimator Moderate or high, if for example: - Indirect adjustment methods - Use of proxy (surrogate) variables - Use of area-level BMI or area-level smoking - BMI as linear term without test of non-linearity - For lung cancer mortality, if smoking is only adjusted as smoking status (never, former, or current), without an additional measure for smoking intensity and/or smoking duration - Insufficient categories of age, iSES/nSES

Subdomain	Low risk criteria (Ideal Study)	Moderate risk criteria	High risk Criteria	Summary of guidance from Traffic Panel to aid interpretation
2.1 Selection of participants into the study (includes entry at baseline and non-response)	Participants in all exposure levels and with all outcomes had equal opportunity to be in the study.	Participants in all exposure levels did not have equal opportunity to be in the study; but not to the extent that effect estimates were seriously biased.	Participants in all exposure levels did not have equal opportunity to be in the study; to the extent that effect estimates were seriously biased.	Low: - All other cases High or moderate if, for example: - Very high enrollment age (i.e. healthy survivor bias), e.g., average baseline age>75 yrs - Selection of a comparison group ("controls") in case-control studies that is not representative of the population that produced the cases (e.g. healthy worker or volunteer effect) - Substantial self-selection or high non-response (>40%) is moderate; High non-response AND likely knowledge of exposure then high (Case-control and cross-sectional) - Differential referral or diagnosis of subjects Note: participants' recruitment designed to maximize exposure contrast should not be considered in this item.
3.1 Methods used for exposure assessment	Exposure levels assessed with appropriate methods.	Exposure levels assessed with less than appropriate methods but not to the extent that effect estimates were seriously biased.	Exposure levels not assessed with appropriate methods to the extent that effect estimates were seriously biased.	Due to the strict exposure assessment framework, this item will be rated low Rob.
3.2 Exposure measurement methods comparable across the range of exposure	Measurement methods used are comparable across the range of exposure.	Measurement methods vary across the range of exposure; however, there is evidence supporting that the exposure measurement is sufficiently similar that effect estimates are not seriously biased.	Measurement methods vary across the range of exposure AND differences are not accounted for.	In air pollution epidemiology studies, this will unlikely occur in practice, and this item will be rated low Rob in almost all studies. Low: - all other cases High - different methods used for high and low exposure, or for cases and controls
3.3 Change in exposure status	Spatial exposure contrasts did not change throughout the study OR time varying exposure appropriately used to account for changes.	Spatial exposure contrasts did change throughout the study AND were not accounted for appropriately, but effect estimates not seriously biased.	Spatial exposure contrasts did change throughout the study AND were not accounted for appropriately AND effect estimates seriously biased and were different in cases and non-cases.	In this item, temporal stability of the exposure model and residential mobility will be considered. Low: - all other cases Moderate: - Concerns with temporal stability of spatial exposure pattern without additional support due to long time span (>5-10 years) - Concerns about high number of relocation without additional support due to long time span. - For birth outcomes, if not accounted for residential mobility during pregnancy High: - See description moderate, but for longer time span (>10 years) - Important local changes of spatial patterns (i.e., newly instituted restricted access zones without support) - Concerns about Quitting ill (cross-sectional and case-control studies)

Subdomain	Low risk criteria (Ideal Study)	Moderate risk criteria	High risk Criteria	Summary of guidance from Traffic Panel to aid interpretation
4.1 Blinding of outcome measurement	Outcome measures were not influenced by knowledge of the exposure.	Outcome measures were influenced by knowledge of the exposure; however, evidence supports that effect estimates were unlikely biased.	Outcome detection was related to exposure status and effect estimates are likely biased.	Low: - All other cases Moderate: - Self-reported outcome without support of additional (objective) measures High: - Self-reported outcome AND likely knowledge of exposure (i.e., traffic indicators) - Examiner not blinded to exposure status
4.2 Validity of outcome measurements	No systematic errors in the measurement of the outcome OR systematic errors were unrelated to the exposure.	Minimum systematic errors suspected in the measurement were related to the exposure received.	Critical systematic errors in the measurement were related to the exposure received.	Low: - Death registry - Administrative database (including hospital admission databases) - Medical records - Controlled exams using standardized procedures and methods - Birth registry data - Validated questionnaire data Moderate: - Self-reported outcome with a non-validated questionnaire - Not-validated controlled exams using not-standardized procedures and methods High: - Moderate AND likely knowledge of exposure (i.e., traffic indicators)
4.3 Outcome measurement	Methods of outcome assessment were comparable across exposure groups	Methods of outcome assessment were not comparable across exposure groups; however, evidence supports that outcome detection would not have varied.	Methods of outcome assessment were not comparable across exposure groups.	Rate this question the same as 4.2 Validity of outcomes.

Subdomain	Low risk criteria (Ideal Study)	Moderate risk criteria	High risk Criteria	Summary of guidance from Traffic Panel to aid interpretation
5.1 Missing data on outcome measures	No missing outcome data OR missing data infrequent (<10%) OR missing data not related to outcome/exposure OR data imputed using appropriate methods OR comparison of complete case and full population effect estimates in scenario analyses suggest no bias.	Missing data on outcomes not infrequent (≥10%) AND rationale for attrition explained in the study; methods have possibly been used to properly account for it.	Evidence of substantial missing outcome data (≥10%), rationale for attrition not explained in the study AND methods unlikely to properly account for it.	Reasons: loss to follow-up, missing appointments, exclusions: Low - All study designs: Missing data/participants < 10-20%. - All study designs: Missing data/participants >20% but with support of no relation to exposure - Worst/best case scenarios suggest no substantial change - For mortality outcomes based on registry data, if no (or very little) information on missing data is given assume missing data is no issue Moderate - Missing data not infrequent (see numbers above) AND appropriate rationale for missingness given - For morbidity outcomes, if no (or very little) information on missing data is given High - Missing data not infrequent (see numbers above), no rationale, no sensitivity analyses
5.2 Missing data on exposures	No missing exposure data OR missing data infrequent (<10%) OR missing data not related to outcome/exposure OR data imputed using appropriate methods OR comparison of complete case and full population effect estimates in scenario analyses suggest no bias.	Missing data on exposure not infrequent (≥10%) AND rationale for attrition explained in the study; methods have possibly been used to properly account for it	Evidence of substantial missing exposure data (≥10%), rationale for missing data not explained in the study, AND/OR the portion of participants and reasons for missing data are dissimilar across exposures/exposure groups.	Same as 5.1 Note: whether exposure data are classified as missing depends on the study definition. It was decided to define missing exposure data as defined in the individual papers, even if that leads to inconsistency. Note: this question is independent from Question 5.1 with one exception: if we can't differentiate whether the missing data had to do with missing exposure or outcomes, then we would rate the two items identically.
6. Authors reported a priori primary and secondary study aims	Effect estimates presented for all hypotheses tested as per aims; reference to published or unpublished study protocol.	Effect estimates presented for some (not all) hypotheses tested as per aims, but evidence suggests that effect estimates unlikely to be seriously biased.	Effect estimates selectively presented for some (not all) hypotheses tested as per aims and effect estimates likely to be seriously biased.	Low: - All research questions or hypotheses from the introduction are addressed Moderate: - Not all research questions of hypotheses from the introduction are addressed High: - Only subgroup analyses from a larger study population presented without reference to an earlier publication or marginally presented in the paper

*The Panel decided to use the risk of bias Tool and Guidance used in the WHO Air Quality Guidelines (AQG) review because the tool was designed for assessment of risk of bias in observational air pollution epidemiology studies (WHO 2020). The tool was modified based upon Panel members' expert judgement and experience in applying the tool in the systematic reviews of the WHO AQG (Chen and Hoek 2020; Huangfu and Atkinson 2020).

References

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; doi:10.1016/j.envint.2020.105974.

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; doi:10.1016/j.envint.2020.105998.

WHO. 2020. RoB Assessment Instrument for Systematic Reviews Informing WHO Global Air Quality Guidelines (2020). Available: *https://www.euro.who.int/en/health-topics/environment-and-health/air-quality/publications/2020/risk-of-bias-assessment-instrument-for-systematic-reviews-informing-who-global-air-quality-guidelines-2020* [accessed 21 September 2020].

5.3 Overall assessment of the epidemiological evidence – further elaborations

5.3.1 Introduction

This document outlines the steps and approaches that the Panel has followed to assess the level of confidence in the evidence that traffic-related air pollution (TRAP) is associated with the selected health effects. This assessment was based on all studies identified in the systematic search – thus studies that entered a meta-analysis as well as the studies that were not used in meta-analysis. The Panel assessed the confidence for a given health outcome by considering the strengths and weaknesses in a collection of human studies that constitute the body of evidence. For this purpose, the Panel decided to follow the methods proposed by the Office of Health Assessment and Translation (OHAT) (OHAT 2019a; Rooney et al. 2014). OHAT serves as an environmental health resource to the public and to health research and regulatory agencies in the US. It conducts technical assessments focused on understanding the potential for adverse effects on human health by agents, substances, mixtures, or exposure circumstances. These evaluations can lead to National Toxicology Program opinions on whether these substances may be of concern given what is known about current human exposure levels.

The OHAT method is based on the methods of Grading of Recommendations Assessment, Development and Evaluation (GRADE), which has been adopted by Cochrane, and many other organizations (Schünemann et al. 2013). OHAT has extended the GRADE approach to include observational human studies in addition to randomized controlled trials. Moreover, OHAT applies the framework separately for animal and human data, which is relevant for the focus of this review on epidemiological studies. The overall OHAT process of rating the confidence in the body of evidence and then translating confidence rating into level of evidence of health effects is summarized below.

The Panel recognized however that the scientific judgments involved in developing these ratings are inherently subjective. A key advantage of the evaluation approach is that it provided a framework to systematically document and explain the decisions made, and thereby provide transparency into the scientific basis of judgments made in reaching conclusions. On the other hand, despite the ongoing attempts to apply the GRADE approach to environmental health (Morgan et al. 2019), the application of the GRADE methods, in particular the risk of bias tools, have been heavily criticized (Bero et al. 2018; Savitz et al. 2019; Steenland et al. 2020). If not carefully applied, the use of those tools and frameworks can become a mechanical exercise that may lead to erroneous conclusions, because the assessments may sometimes consider individual studies out of context, and do not take a broader approach of the evidence.

The Panel noted several challenges in applying the OHAT methods in its original form in the current review. A major issue is the initial level of confidence assigned to observational studies. Typically, GRADE and OHAT consider randomized controlled trials as the gold standard for judging observational studies in environmental epidemiology and therefore epidemiologic studies have a lower initial confidence. This approach originates from clinical medicine to evaluate treatments and objectively distinguish effective from ineffective ones and places a high priority on avoiding false positive conclusions (e.g., recommending treatments that do not work). This leads to a hierarchy of types of evidence that puts randomized controlled trials at the top. In environmental epidemiology, the evidence rarely comes from randomized controlled trials and, rather than avoiding false positives, the greater concern is avoiding false negatives (e.g., failing to detect a specified hazard). Each study design is a proxy of some inherent strengths and weaknesses. Thus, when applying GRADE and OHAT, studies may be “penalized” twice for the same issue, such as a lack of randomization of exposure and possibility of residual confounding. Randomized controlled trials are largely infeasible in environmental epidemiology, as one cannot ethically randomize people to potentially harmful exposures. Beyond that, randomized controlled trials typically involve limited sample sizes and a short follow-up time, which is often inadequate for observing chronic disease or rare outcomes. Also, randomized controlled trials deliver the exposure (e.g., medication) at the beginning of follow-up,

typically in a limited number of dose levels which does not mimic real-life circumstances. Moreover, randomized controlled trials may involve highly selective study groups meeting particular criteria, which may have little generalizability to other populations (Steenland et al. 2020). An important strength of large epidemiological study is the inclusion of the full spread of susceptibility – not typically met in randomized controlled trials.

Therefore, the Panel have used the OHAT method as a guide and did not apply the methods in a mechanistic way. Some features of the OHAT methodology remain controversial. For example, some heterogeneity is expected across studies due to the nature of observational studies in different populations, contexts, and exposure conditions, and does not necessarily reduce confidence in the body of evidence based on inconsistency. Hence, the Panel have slightly modified the OHAT approach to better fit the needs of the Panel, as summarized below. The changes were also based on the NTP Monograph on Systematic Review of Traffic-Related Air pollution and hypertensive disorders of pregnancy (OHAT, 2019b) as well as on the suggestions from the recent COSMOS-E: Guidance on conducting systematic reviews and meta-analyses of observational studies of etiology (Dekkers et al. 2019).

Despite modifications, the Panel was convinced that the OHAT methods remains imperfect, and its application was challenging. The Panel thought the application of the OHAT methods was most useful to evaluate *the quality of the body of evidence* of studies entering a meta-analysis – irrespective of the strength and nature of the association. The Panel thought it was prudent to accompany the OHAT assessment with a broader approach to assess *the level of confidence in the presence of an association*, considering the meta-analysed studies as well as other studies not entering the meta-analysis. Note that the goal of the overall evaluation is to establish the collective assessment of confidence in the presence of an association, not of the exact magnitude of the effect estimate. To this end, the Panel also took a broader approach and developed a narrative assessment for each heath outcome. This additional assessment considered in more detail the populations studied, the size of the evidence base, the results of the meta-analyses and of the studies not entering any meta-analyses, the consistency of the results for single pollutants and across pollutants and indirect traffic measures, and other considerations. In other words, the emphasis of the narrative assessment is on the overall results and their interpretation, while taking into account the validity issues related to the study design (e.g., confounding, selection bias, chance).

The narrative assessment and the assessment based on the modified OHAT approach were considered complementary, reflecting the complex issues in determining the level of confidence.

Below a summary of both methods is given, including the main differences, as well as the main modifications of the OHAT methods to better fit the needs of the Panel. Figure 1 gives a summary of the overall approach taken in the traffic review.

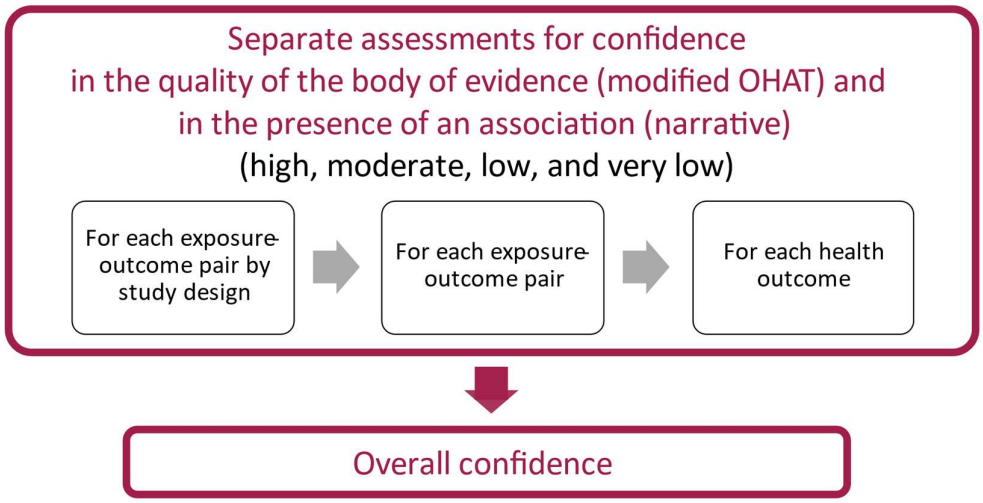

Figure 1. Summary of overall confidence assessment for TRAP and selected health outcomes

5.3.2 Summary of the OHAT method

OHAT considers as the body of evidence the studies whose results can be summarized in a meta-analysis as well as the studies that will not lend themselves to meta-analysis. While this may be true, the Panel noted that the framework is heavily geared towards the studies entering a meta-analysis. The OHAT method uses four descriptors to indicate the level of confidence in a body of evidence, see Table 1 and Figure 2, which are also included in the traffic review protocol (HEI 2019).

Table 1. Confidence ratings in the body of evidence (OHAT 2019a; Rooney et al. 2014).

Confidence rating	Definition
High confidence (++++)	High confidence in the association between exposure to the substance and the outcome. The true effect is highly likely to be reflected in the apparent relationship.
Moderate confidence (+++)	Moderate confidence in the association between exposure to the substance and the outcome. The true effect may be reflected in the apparent relationship.
Low confidence (++)	Low confidence in the association between exposure to the substance and the outcome. The true effect may be different from the apparent relationship.
Very low confidence (+)	Very low confidence in the association between exposure to the substance and the outcome. The true effect is highly likely to be different from the apparent relationship.

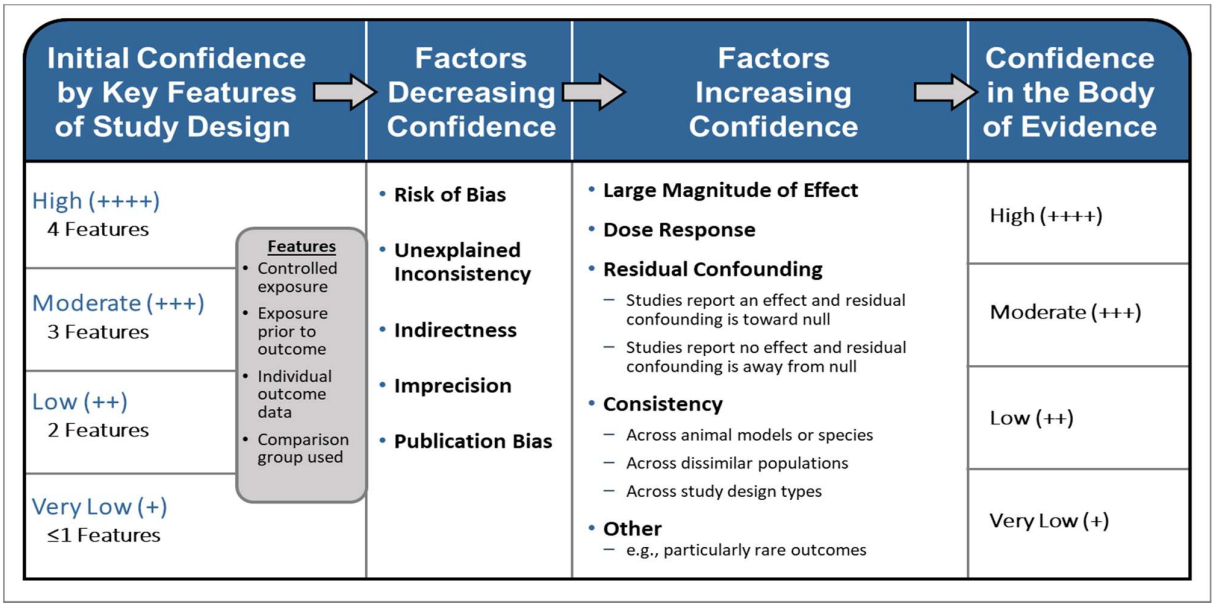

Figure 2. Assessing confidence in the body of evidence (OHAT 2019a).

Available studies on a particular outcome or health endpoint are initially grouped by key study design features, and each grouping of studies is given an initial confidence rating by those features. This initial confidence rating (column 1) for the body of evidence from this group of studies is downgraded for factors that decrease confidence in the body of evidence (risk of bias, unexplained inconsistency, indirectness or lack of applicability of the body of evidence, imprecision of the estimates, and publication bias) and

upgraded for factors that increase confidence in the body of evidence (large magnitude of effect, monotonic exposure response, consistency across study designs/populations/animal models or species, and consideration of residual confounding or other factors that increase the confidence in the association). If a decision to downgrade is borderline for two factors, the body of evidence is downgraded once (for a single factor) to account for both partial concerns based on considering the key drivers of the strengths or weaknesses. Similarly, the body of evidence is not downgraded twice for what is essentially the same limitation (or upgraded twice for the same asset) that could be considered applicable to more than one factor of the body of evidence. Consideration of consistency across study designs, human populations, or animal species is not included in the GRADE guidance; however, it is considered in the modified version of GRADE used by OHAT (Rooney et al. 2014).

In OHAT, the four key study design features used to delineate groups of studies for initial confidence ratings are: (1) the exposure is experimentally controlled; (2) the exposure assessment demonstrates that exposures occurred prior to the development of the outcome (or concurrent with aggravation or amplification of an existing condition); (3) the outcome is assessed on the individual level (i.e., not through population aggregate data), and (4) an appropriate comparison group is included in the study. The first key feature, controlled exposure, reflects the ability of experimental studies to largely eliminate confounding (on average) by randomizing allocation of exposure. Therefore, experimental studies usually have all four features and receive an initial rating of high confidence. Observational studies do not allow for controlled exposure and are differentiated by the presence or absence of the three remaining study design features. For example, cohort studies usually have all three remaining features and receive an initial rating of moderate confidence.

As a final step, to translate confidence ratings into level of evidence for health effects in OHAT (See Figure 3 and Table 2), the nature of the association (health effect or no health effect) is considered. Three descriptors (high, moderate, and low level of evidence) directly translate from the ratings of confidence in that the exposure is associated with a heath effect. If the confidence rating conclusion is very low or no evidence is identified, then the level–of–evidence conclusion is characterized as inadequate evidence. The descriptor evidence of no health effect is used to indicate confidence that the exposure is not associated with a health effect. Because of the inherent difficulty in proving a negative, the conclusion evidence of no health effect is only reached when there is high confidence in the body of evidence.

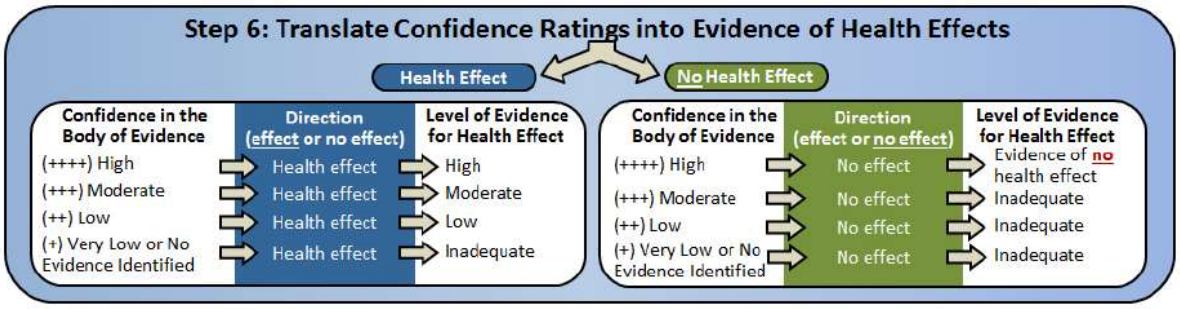

Figure 3. Translate confidence ratings into evidence of health effect conclusions (OHAT 2019a).

Table 2. Level of evidence ratings for health effects (OHAT 2019a; Rooney et al. 2014).

Evidence descriptors	Definition
High level of evidence	There is high confidence in the body of evidence for an association between exposure to the substance and the health outcome(s).
Moderate level of evidence	There is moderate confidence in the body of evidence for an association between exposure to the substance and the health outcome(s).
Low level of evidence	There is low confidence in the body of evidence for an association between exposure to the substance and the health outcome(s), or no data are available.
Evidence of no health effect	There is high confidence in the body of evidence that exposure to the substance is not associated with the health outcome(s).
Inadequate evidence	There is insufficient evidence available to assess if the exposure to the substance is associated with the health outcome(s).

5.3.3 Adaptation of the OHAT method for confidence assessment of the body of evidence for the traffic review

The text below elaborated on the main modifications of the OHAT method to better fit the need of the traffic review.

Before entering into the details, it should be noted that OHAT automatically translates confidence ratings in the body of evidence into level of evidence for health effects where it considers the nature of the association (health effects or no health effect) (Figure 3 and Table 2). The Panel was convinced that this automatic translation was not appropriate as it transferred the confidence in the body of evidence (mainly the results of the evaluation of the quality of the studies) into an evaluation of the level of the evidence, without considerations of additional relevant factors, such as strength and nature of the association and the consistency of the results from the meta-analyses and the studies not meta-analyzed. Thus, also given the charge of the Panel, the Panel focused on a statement about the confidence in the body of evidence as high, moderate, low and very low. The Panel noted that convincingly demonstrating no health effect is generally beyond what epidemiological studies can achieve.

Another important choice was whether the downgrading and upgrading of the confidence based on the factors listed above are 'independent', i.e., an upgrade can occur if the confidence has been downgraded for other factors. The Panel made the choice to evaluate independently the downgrading and upgrading factors without imposing a constraint, following the procedures applied in the WHO systematic reviews of air pollution and traffic noise (Chen and Hoek 2020; Huangfu and Atkinson 2020; WHO 2018).

Because TRAP is a complex mixture, the Panel decided to evaluate the body of evidence separately for each exposure metric included in the review (Step 1, confidence for individual TRAP components), and then evaluate the body of evidence for the health effects across all included traffic–related air pollutants and indirect traffic measures (Step 2, confidence for combined measures of TRAP). Thus, confidence rating for each health endpoint was first developed separately for each exposure metric. Then, the confidence in the body of evidence was considered for the combined TRAP exposure.

The OHAT confidence rating is heavily geared towards the studies entering a meta-analysis. The Panel did not apply the confidence assessment for the exposure-outcome pairs if no meta-analysis was conducted due to few studies. The Panel also did not conduct meta-analysis of studies based on indirect traffic measures, such as distance or traffic density variables, due to limited comparability across studies.

The results from studies that did not enter a meta-analysis were mainly considered in the narrative assessment (see below). However, the modified OHAT assessment mentioned those studies, in particular when they were large and informative, to inform the overall evaluation across study designs and different pollutants (Step 2), but only if Step 1 could be completed, meaning that at least one meta-analysis was conducted for an outcome.

5.3.3.1 Confidence for individual TRAP components (Step 1)

The aim of this step was to assess the level of confidence for each individual exposure-outcome pair for two groups of study designs separately (cohort and case-control studies considered together as one group; cross-sectional studies considered as a separate group). The separation of studies is justified because the initial starting point is "moderate" for both cohort and case–control studies and "low" for cross–sectional studies. Table 3 was prepared for each outcome and a final integration of evidence over both groups of study designs was made. The table contained information as indicated in the original OHAT document (Table 4). The outcome of this Step 1 was one confidence rating per exposure-outcome pair based on the highest confidence level reached for each study design. The separation of study design broadly agreed with the decision to conduct separate meta-analyses for incidence (typically cohort studies) and prevalence (typically cross-sectional) studies, where relevant, though the reality was more complex, at least for some respiratory outcomes. For example, some birth cohorts assessed the occurrence of a specific respiratory condition at specific ages during the course of the follow-up and the results were properly analyzed as prevalence (at a specific age) rather than incidence. In this case, the design of the study was a cohort although the specific analysis was cross-sectional. Also, regarding the outcome ALRI, the Panel considered all ALRI studies as incidence studies given the acute nature and expected absence of the infection prior to diagnosis and/or between repeated infections in the same individual. Hence for this outcome, also because there were few cross-sectional studies available, different study designs were combined, and the subsequent confidence assessment was not separately assessed by study design. Additional sensitivity analysis by study design was conducted.

The Panel agreed that each assessment needed at least three studies in the same study design group, thus three cohorts /case-control studies or three cross-sectional studies. In cases of fewer studies, the Panel described those studies and evaluated narratively how they support or did not support the evaluation done on the basis of the other pollutants.

5.3.3.2 Initial rating of the body of evidence based on the study design

Observational studies cannot assign controlled exposures as requested in the OHAT protocol, therefore prospective cohorts receive the initial rating of "moderate confidence" as they have all the three other conditions illustrated in Figure 1 (i.e., exposure precedes the outcome, individual data, and comparison group). OHAT further proposes an initial rating of low to moderate for case-control studies and an initial rating of low for cross-sectional studies.

The Panel decided to follow the same logic and considered cohort studies (both retrospective and prospective) and case-control studies (based on incidence outcome measures) to have an initial rating of moderate confidence because the other three key features used for the initial confidence assessment were often met (exposure precedes the outcome, individual-level data, comparison group). The Panel noted that potential issues related to the internal validity of both retrospective and prospective cohorts would be covered in the risk of bias assessment and they thought there were no reasons to differentiate them for the initial rating.

Similar to OHAT, the Panel decided to start with an initial rating of low confidence for cross-sectional studies because one cannot typically assert that the exposure precedes the outcome. Ecologic studies and case reports were excluded from the traffic review and therefore no initial rating was required.

Table 3. Confidence rating for traffic-related air pollution and a specific outcome (only for the pollutants where a meta-analysis was conducted).

	High ++++ Moderate +++ Low ++ Very low +		**Factors decreasing confidence "0" if no concern; "-" if serious concern to downgrade confidence**					**Factors increasing confidence "0" if not present; "+" if sufficient to upgrade confidence**					
Pollutant	**Study design**	**Initial confidence rating** (# studies)	Risk of Bias	Unexplained inconsistency	Indirectness*	Imprecision	Publication bias	Magnitude of the effect*	Monotonic exposure-response	Consideration of residual confounding	Consistency across populations	**Confidence rating by study design**	**Final confidence rating**
NO_2	Cohort / case-control												
	Cross-sectional												
NO_X	Cohort / case-control												
	Cross-sectional												
NO	Cohort / case-control												
	Cross-sectional												
CO	Cohort / case-control												
	Cross-sectional												
$PM_{2.5}$	Cohort / case-control												
	Cross-sectional												
PM_{10}	Cohort / case-control												
	Cross-sectional												
EC, BC, BS, soot	Cohort / case-control												
	Cross-sectional												
Other pollutants	Cohort / case-control												
	Cross-sectional												
Combined measures of TRAP	Cohort / case-control												
	Cross-sectional												

*Was not used in the traffic review – see explanation below. Cells in gray are not applicable.

Table 4. Evidence profile table format (from OHAT 2019a).

Factors decreasing confidence					**Factors increasing confidence**			
Risk of Bias	Unexplained inconsistency	Indirectness*	Imprecision	Publication bias	Magnitude of the effect*	Monotonic exposure-response	Consideration of residual confounding	Consistency across populations
Serious or not serious	Serious or not serious	Serious or not serious	Serious or not serious	Detected or undetected	Large or not large	Yes or No	Yes or No	Yes or No
-Describe trend -Describe key questions -Describe issues	-Describe results in terms of consistency -Explain apparent inconsistencies (if it can be explained)	- Discuss use of upstream indicators or populations with less relevance	- Discuss ability to distinguish treatment from control -Describe confidence intervals	- Discuss factors that might indicate publication bias (e.g., funding, lag)	- Describe magnitude of response	- Outline evidence for or against exposure-response	- Address whether there is evidence that confounding would bias toward null	- Describe cross-species, model, or population consistency

*Was not used in the traffic review – see explanation below.

5.3.3.3 Factors that may reduce confidence

On an outcome-by-outcome basis, five properties for a body of evidence (risk of bias across studies, unexplained inconsistency, indirectness in the body of evidence, imprecision of the estimates, and publication bias) were used to determine if the initial confidence rating based upon study design features should be downgraded.

1. Risk of Bias

Risk of bias for a given outcome was considered for all studies included in the meta-analyses. To this end, the Panel used the Risk of Bias Tool and Guidance from the WHO AQG review (WHO, 2020). The tool was modified based upon Panel members' expert judgement and experience in applying the tool in the systematic reviews of the WHO AQG (Chen and Hoek 2020; Huangfu and Atkinson 2020). The risk of bias assessment was conducted for each exposure–outcome pair.

The Panel prepared a visual summary of the risk of bias ratings for each study and each exposure-outcome pair that entered the meta-analyses (see example table 5 below). The Panel also included the risk of bias rating tables showing the individual study ratings as an appendix.

The risk of bias assessment provided an overview of the specific forms of bias for all studies included in the meta-analyses. In addition, it highlighted specific threats to validity that could be explored when evaluating inconsistency within the evidence base.

OHAT suggests that the decision to downgrade a group of studies because of risk of bias should be applied conservatively and be reserved for groups of studies for which there is substantial risk of bias across most of the included studies composing the body of evidence and/or for those studies that have the most weight in the meta-analysis. The Panel performed sensitivity analyses comparing effect estimates obtained from studies of high risk of bias and low/moderate risk of bias per domain of the risk of bias tool, in case at least three studies were rated as high risk of bias. Note that these sensitivity analyses were conducted per bias domain (e.g., exposure assessment, confounding). No summary classification was derived across the six domains.

These sensitivity analyses were crucial because the risk of bias tool only assessed whether there is a potential for bias not whether there is an actual bias. When effect estimates from studies with low/moderate and high risk of bias were virtually the same, the Panel did not downgrade the evidence and include all studies in the overall assessment. When effect estimates from studies with low/moderate were considerably different from estimates of studies at high risk of bias and there were sufficient studies in the low/moderate categories, the high risk of bias studies were omitted from the confidence-rating phase entirely, and the Panel focused on studies at low and moderate risk of bias only, irrespective of the direction of the difference. In such case, there was no reason to downgrade because the body of evidence on which the conclusions was based considered studies of low and moderate risk of bias only. Downgrade occurred only when effect estimates from studies at low/moderate were considerably different from estimates of studies at high risk of bias and the body of the evidence with low/moderate risk of bias was limited (few studies and/or small weight in the meta-analysis). In summary, the decision to downgrade because of risk of bias was applied after careful review of the visual summary of the risk of bias tool for each exposure-outcome pair (Table 5) and the results of the sensitivity analyses.

Table 5. Summary of risk of bias rating for studies on a specific outcome.

		Per study			Per pollutant-study pair		
Domain	**Subdomain**	**Low-risk**	**Moderate -risk**	**High-risk**	**Low-risk**	**Moderate -risk**	**High-risk**
1.Confounding	Were all important potential confounders adjusted for in the design or analysis?						
	Validity of measuring of confounding factors						
	Control in analysis						
	Overall						
2.Selection Bias	Selection of participants into the study						
3.Exposure assessment	Methods used for exposure assessment						
	Exposure measurement methods comparable across the range of exposure						
	Change in exposure status						
	Overall						
4.Outcome measurements	Blinding of outcome measurements						
	Validity of outcome measurements						
	Outcome measurements						
	Overall						
5.Missing data	Missing data on outcome measures						
	Missing data on exposures						
	Overall						
6.Selective reporting	Authors reported a priori primary and secondary study aims						

2. Unexplained inconsistency

Large variability in the magnitude and direction of individual study effect estimates can reduce confidence in the body of evidence. However, there are several legitimate reasons that may plausibly account for variability in magnitude of effect estimates, including different populations, exposure assessment methods, pollution mixtures or co-pollutants, time period, age structures, and follow-up time across studies. Additionally, a non-linear relationship between the exposure and the outcome could be responsible for the heterogeneity of the effect estimates across studies if the different populations have different average exposure levels, clearly demonstrated in the analysis of short-term effects of $PM_{2.5}$ on mortality in 652 cities worldwide (Liu et al. 2018). This is commonly seen as effect modification by level of exposure. In the traffic review protocol, the Panel have identified subgroups of interest for potential sensitivity analyses, provided there were sufficient studies, such as geographical areas, time period, high risk of bias versus lower risk of bias per domain of the Risk of Bias Tool, and confounder adjustment for individual–level behavioral factors. The statistical power of studies was also considered if the Panel detected an inconsistency of findings across studies.

Note that no single statistical measure of consistency of findings across studies is ideal, and the following factors were considered when determining whether to downgrade for inconsistency: (1) direction and magnitude of the point estimates, (2) extent of overlap between confidence intervals, and (3) results of statistical measures of heterogeneity, e.g., Cochran's Q (chi-square, χ^2), I^2, τ^2 (tau-squared), and prediction intervals. There are well known limitations of statistical tests for heterogeneity, and they are less reliable when there are only a few studies. Given these limitations, the Panel decided to primarily interpret I^2, where I^2 values of <50% were interpreted as low; between 50 and 75 as moderate; and then >75 as high degree of heterogeneity, respectively (Woodward 2013). Note that thresholds for the interpretation of I^2 can be misleading, since its value also depends on the magnitude, direction and precision of the effect estimates from the individual studies (Rücker et al. 2008). OHAT provides slightly different thresholds, e.g., between 50 and 90 as substantial; and 75 to 100 as considerable heterogeneity (OHAT 2019a). This distinction was considered less useful by the Panel because the thresholds are not mutually exclusive, reflecting the challenges of thresholds for the interpretation of I^2.

The decision to downgrade because of unexplained inconsistency was considered if heterogeneity was high and applied after careful review of the potential sources of heterogeneity, including risk of bias, and considering the direction of the effect estimate rather than its magnitude. It is worth pointing out that inconsistency in the magnitude of an association was much less of an issue compared to inconsistency in direction. A group of studies all showing an association, albeit with inconsistent magnitude, was of much less concern to the Panel as the purpose of the assessment was identification of the presence of an association rather than estimation of the exact magnitude of the association.

3. Indirectness

This criterion refers to the applicability, external validity, generalizability, and relevance of the studies in the evidence base in addressing the objectives of the evaluation. Directness addresses the question, "Did the study design address the topic of the evaluation?" The OHAT handbook also considers the appropriateness of the window of exposure given the health outcome measured as part of the evaluation for directness and applicability as well as the duration of follow-up. Because of the selection of human studies, exposure specificity, and the selection of the outcome measures in the traffic review, this factor will not lead to a downgrade in practice and was not considered further.

4. Imprecision

Precision reflects the degree of certainty surrounding an effect estimate with respect to a given outcome (AHRQ, 2013), and in single studies depends on size of the study and the magnitude of the association, among others. This relates to individual studies and does not readily translate into a random effect meta-analysis. In random effects meta-analysis the issue is complicated by the possible heterogeneity of the association across studies. This is because the confidence interval of a random effects estimate depends upon the precision of the individual study estimates and the t^2 (and hence the number of studies). Therefore, one could be satisfied about the power in individual studies but still have imprecision in the random effect summary estimate if t^2 is large.

The issue of imprecision is also related, and limited, to the specific purpose of the systematic review, hazard identification vs quantitative risk assessment (Samet et al. 2020; Saracci, 2017). In the hazard identification process, as in the Traffic Review, the goal of the overall evaluation is to establish the confidence in the presence of an association and the interest is whether the overall effect estimate departs from the null. In a quantitative risk assessment, the interest is in the exact magnitude of the association and its precision, as in the WHO AQG reviews.

Therefore, in the assessment of imprecision, the Panel decided to adapt the GRADE approach that combines multiple parameters: power (under specified alpha and beta's), width of the confidence intervals of the random effect meta-analysis, and specified critical margins of "no effect", "important benefit", or

"important harm". The Grade Handbook prescribes simple rules: if the optimal information size criterion is not met, rate down for imprecision; if the optimal information size criterion is met and the 95% confidence interval excludes unity, do not rate down for imprecision; if the optimal information size criterion is met, and the 95% confidence interval includes unity, rate down for imprecision as the confidence interval fails to exclude important benefit or important harm. It should be recognized that the application of such rules has been challenging even in the context of clinical medicine (Castellini et al. 2018).

In its assessment of imprecision, the Panel considered thus the number of the participants included in the meta-analysis and the width of the 95% confidence intervals if interval clearly included unity. The rules were defined as follows (see also Figure 4):

-If the total number of participants included in the systematic review was less than the number of participants generated by a conventional sample size calculation for an individually powered study, the Panel downgraded for imprecision if the 95% confidence interval included unity. In the (unlikely) event that the power was not sufficient, but the 95% confidence interval excluded unity, the Panel did not to downgrade for imprecision.

-If the criterion for study power was met and the 95% confidence interval excluded unity (regardless of width, and allowing for some flexibility in case of marginally overlapping confidence intervals, the Panel did not downgrade for imprecision.

-If the criterion for study power was met, and the effect estimate was precise with a narrow 95% confidence interval, and the confidence interval included unity, the Panel did not downgrade for imprecision. For example, fairly precise effect estimates indicating a null or negative association (RR close to 1) were not downgraded by the Panel. Note that the presence of the association, i.e., whether there is an association or not, was included in the narrative assessment.

-If the criterion for study power was met, but the effect estimate was imprecise with a wide 95% confidence interval and the confidence interval clearly included unity, the Panel downgraded for imprecision.

For ratio measures (like RR), a narrow (precise) confidence interval was defined as a difference on the log scale ≤0.1 from the upper to the lower 95% confidence limit. A wide (imprecise) confidence interval was thus defined as a difference on a log scale >0.1 between the upper and lower 95% confidence limits (Rothman and Greenland 2018; Zhang et al. 2019).

For the single difference measure included in the traffic review (term birth weight), a narrow confidence interval was defined as 0.1 times the expected standard deviation of birth weight of a full-term new-born, that could be approximated as 400 grams. Hence if the difference between the limits of the CI was <=40 grams this was considered as narrow, otherwise wide. The choice of the selected width is based on clinical considerations of the outcome, as there is no prior literature on interpreting precision in meta-analytic estimates of continuous outcomes.

Furthermore, to calculate statistical power, a simple calculator for prevalence studies or for incidence studies and a dichotomous exposure was used: https://www.stat.ubc.ca/~rollin/stats/ssize/b2.html. For example, the following tables provided some general indications about the sample size needed for a RR (or OR) of 1.05, 1.10, and 1.20 (minimum effect size) that can be applied to the morbidity studies. For mortality studies, a lower minimum effect size was chosen, e.g., RR=1.02.

Table 6. Total sample size needed for prevalence studies or incidence studies using a dichotomous exposure [1].

Incidence (% in low exposed group)	RR=1.05	RR =1.10	RR=1.20
5	122,124	31,234	8,158
10	57,763	14,751	3,841
15	36,310	9,257	2,402
20	25,583	6,510	1,683

[1]assuming enrollment ratio 1:1 (exposed: unexposed), RR observed in the meta-analysis and a comparison below versus above the median of the pollutant level (alpha=0.05, Power 0.80).

Table 7. Total sample size for case-control studies using a dichotomous exposure[1].

% Exposed in the control group	RR =1.05	RR =1.10	RR =1.20
50	26,383	6,918	1,895

[1]assuming a case/control ratio 1:1 and a comparison below versus above the median of the pollutant level (alpha=0.05, Power 0.80).

Finally, as imprecision may be the result of heterogeneity ad described above, often it is difficult to distinguish between wide confidence intervals of the meta-analytic estimates due to heterogeneity versus those due to imprecision, which leads to the question of whether to downgrade once or twice. In most cases, a single downgrade for one of these factors is sufficient. Thus, in most cases where the body of evidence was downgraded for inconsistency, the Panel did not further downgrade for imprecision. However, it was considered appropriate to downgrade twice if studies were both very inconsistent and imprecise. Similar to OHAT, the Panel included the possibility to omit statistically underpowered studies from consideration entirely when determining confidence ratings, depending on the number of studies.

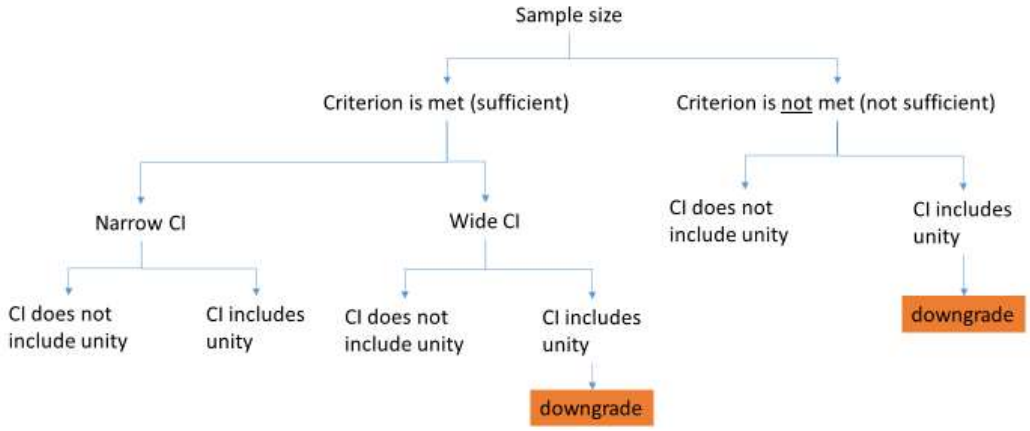

Figure 4. Flowchart of the assessment of imprecision in the traffic review.

5. Publication Bias

The factor publication bias refers solely to the evaluation to the *group of studies* in each exposure-outcome and study design category. This factor should not be confused with the Risk of Bias Tool item selective reporting which was assessed at the *individual study level*, and for which potential downgrading of that item will be considered under risk of bias.

OHAT suggests categorizing publication bias as "undetected" (no downgrade) or "strongly suspected". Some degree of publication bias is likely and downgrading is reserved for instances where the concern is serious enough to significantly reduce confidence in the body of evidence as a whole. Funnel plots and Egger's regression tests were used to visualize asymmetrical or symmetrical patterns of study results to help assess publication bias, provided there were sufficient studies. The Panel additionally applied the trim and fill method to determine whether potential publication bias produces a meaningful change in effect estimates (Shi and Lin, 2019). Those methods are recommended when at least 10 studies are included in the meta-analysis. However, even 10 studies may be low, because the results of the Egger tests also depend on study size and magnitude of associations (Lin and Chu 2018; Murad et al. 2018). Furthermore, true heterogeneity of effect size may lead to asymmetrical funnel plots and significant Egger tests. Finally, the interpretation of the funnel plots assumes that small studies with imprecise effect estimates are less expensive than large studies (small study bias). However, in air pollution epidemiology the relationship between study size and cost can be weak because of the extensive use of existing (and often very large) administrative databases in air pollution epidemiology.

Hence all those methods were applied with caution and the possibility to explore publication bias was limited. The OHAT handbook lists additional tools for detecting potential publication bias: tracking of conference abstracts that do not make it into publications within 3-4 years; role of funding source; and suspicion of early positive studies, especially when studies are small. The last approach was also explored in the traffic review. First, a sensitivity analysis was conducted for studies before vs after 2008. Second, the Panel prepared plots of the number of participants versus publication year, coloured by statistical significance of results for all estimates and for only those included in meta-analysis.

5.3.3.4 Factors that may increase confidence

On an outcome-by-outcome basis, three properties for a body of evidence (large magnitude of effect, evidence of a monotonic exposure response function, and little residual confounding) are considered in both the OHAT protocol and the GRADE guidance to determine if the confidence rating should be upgraded. OHAT also considers an additional factor to address consistency across study designs, human populations, or animal species.

1. Large magnitude of the effect

The GRADE approach upgrades the quality of the evidence in observational studies if the effect size is large or very large (i.e., large RR > 2 or very large RR > 5) because residual confounding is less likely. These numbers are difficult to be applied in environmental health and OHAT proposes that considerations for identifying a large magnitude of effect are made on a project-specific basis based on discussion by the evaluation team. The discussion should be grounded in consideration of the effect being measured, the background prevalence or rate for that effect, and the exposure pattern in human studies including peaks, magnitude, and duration. In the WHO AQG, the calculation of the E-value (VanderWeele and Ding 2017) was proposed to make a judgement about the likelihood that unmeasured confounding could explain away the pooled effect size resulting from the meta-analysis. The application of the E-value, however, needs some arbitrary judgment on the extent of the relationship between the exposure and the confounder (after adjustment for the measured confounders) and its application has been criticized (e.g., Ioannidis et al. 2019). In the traffic review, the Panel consider a "large" effect to be both ambiguous to define and unlikely to occur. Thus, the Panel has decided not to consider this specific upgrading factor.

2. Monotonic Exposure Response Function

In the traffic review, studies that purposefully evaluated the shape of the exposure–response function were noted. The Panel upgraded the confidence level when there was convincing evidence of a plausible monotonic exposure–response gradient in the exposure range of the majority of the studies. Linear, supralinear and sublinear (possibly with a threshold) functions were all interpreted as plausible monotonic exposure-response functions in this context. No upgrade was applied if there was not convincing evidence across studies of a plausible monotonic exposure–response function. The Panel did not accept a statement of no deviation from linear if the linear association was null.

In meta-analyses, effects are expressed using a standardized increment in exposure, assuming a linear exposure-response relationship. Thus, one can argue that this may be sufficient for an upgrade already, but the Panel decided that at least two influential studies should have evaluated the actual form of the relationship (e.g., using splines or quantile analyses) and document a monotonic exposure-response function.

3. Factors potentially shifting the RR towards the null

Residual confounding refers to consideration of unmeasured determinants of an outcome unaccounted for in an adjusted analysis that are likely to be distributed unequally across exposure groups. If the majority of studies report an association despite the presence of residual confounding, or other factors deemed likely to be acting in the opposite direction, confidence in the body of evidence is increased. According to OHAT, upgrading should be considered when there are indications that residual confounding or other factors are likely to lead to an underestimation of an apparent association (i.e., bias towards the null), or when results suggest a spurious protective effect when factors are at work that most likely lead to a bias towards a protective effect. The Panel carefully reviewed whether there was evidence that sources of bias would bias towards null across the studies.

4. Consistency

OHAT proposes that three types of consistency in the body of evidence can be used to support an increase in confidence in the results: 1) across animal studies– consistent results reported in multiple experimental

animal models or species; 2) across dissimilar populations– consistent results reported across populations (human or animal) that differ in factors such as time, location, and/or exposure; 3) across study designs– consistent results reported from studies with different design features, e.g., between cohort and cross-sectional studies in humans or between chronic and multigenerational animal studies. In our review, only the last two factors are relevant. Note that the factor consistency is typically less relevant to establish confidence in the body of evidence, but it is more important when the confidence in the presence of an association is the main focus, as in the narrative assessment. The Panel considered upgrading for consistency across populations when there was clear evidence of an association across different populations, specifically in different geographical areas and between different time periods. Consistency across study design was considered at the next stage of evidence synthesis.

5.3.3.5 Combined confidence for all study designs

When the final assessment **by study design** was completed, for each specific exposure-outcome pair a combined assessment **across different study designs** was carried out. For some of the exposure-outcome pairs only one study design might be available (e.g., most mortality studies are cohort studies) but for other outcomes the evidence base represented a combination of cohort studies, case-control studies and cross-sectional studies (e.g., diabetes).

The confidence rating of an exposure-outcome pair **across different study designs** was equal to the highest rating for an individual study design. As OHAT suggests, consistency across study designs could be a reason to upgrade a confidence rating, but inconsistency across study designs is not necessarily a reason to downgrade a confidence rating. It should be noted that consistency of the results across study designs is an important aspect that has been recently discussed with the concept of triangulation (Lawlor et al. 2016). The underlying premise is that if different epidemiological approaches, possibly with unrelated sources of bias, all support the same conclusion, the confidence in that conclusion is strengthened. This seems particularly compelling when the key sources of bias of some of the approaches are predicted to influence estimates in opposite directions (Pearce et al. 2019). For these reasons, the Panel upgraded the confidence when the results were based on different study designs (cohort studies/case-control versus cross-sectional studies) supporting the same conclusions. Note that such comparison can either use meta-analysis results or results from individual studies using different study designs. In summary, the final judgment regarding an individual pollutant-outcome pair across study design types will consider the evidence with the highest confidence and then, if necessary, upgrade or downgrade according to the confidence rating of the other study design.

5.3.3.6 Assessment of confidence for combined measures of TRAP (Step 2)

The overall assessment of the body of evidence for TRAP and a selected outcome was based on the final ratings of each exposure-outcome pair from Step 1. The result of Step 2 is a rating of confidence for overall TRAP exposure and association with each specific health outcome. Conclusions for the combined confidence are primarily based on the evidence with the highest confidence of a pollutant. However, such a conclusion was upgraded or downgraded, if needed, on the basis of the confidence rating of the other pollutants, information from large and informative studies not entering a meta-analysis, as well as on the basis of traffic specificity. For instance, if high confidence was provided for only one pollutant, the panel may decide whether the overall evidence is high or moderate depending on the assessment for the other pollutants.

The exposure framework developed for the traffic review specifies the general exposure criteria for use in selection of epidemiological studies to be included in the review (see Chapter 6 for details). Additional (stricter) criteria were developed to identify studies where there was high confidence that the exposure contrasts in the study were because of traffic as evidenced by their high spatial resolution and exposure

assessment methods that would capture variation in traffic-related air pollutant concentrations (high traffic specificity).

To decide on whether a downgrade or upgrade was warranted, the Panel performed sensitivity analyses comparing effect estimates obtained from studies of high traffic specificity and the other studies in cases where at least three studies were rated as high traffic specificity. When effect estimates for a specific outcome from studies with high traffic specificity versus other studies were similar, the confidence in the body of evidence remained the same. When effect estimates from studies with high traffic specificity versus other studies reported a larger magnitude of effect, the Panel upgraded the evidence. Please see Table 8 summarizing these general decisions.

As with all the factors listed above, downgrading and/or upgrading based on traffic specificity should reflect the entire body of studies; therefore, this decision was applied conservatively and reserved for cases for which there is substantial evidence of traffic specificity (or lack thereof) across most of the studies composing the body of evidence and/or for those that have the most weight in the meta-analysis.

Table 8. General decisions when comparing results with high traffic specificity versus other studies.

Comparison of results with high traffic specificity versus other studies	**Decision**
Stronger evidence	Upgrade
Comparable evidence	No change
Weaker evidence	Downgrade
Opposite evidence	Downgrade
Insufficient number of studies	No change

5.3.4 Narrative assessment of the level of confidence in the presence of an association

There is large and consolidated tradition in evidence synthesis and integration in environmental health sciences to arrive at an overall assesment of the strength of the evidence. The tradition has its basic principles in the application on the Bradford-Hill criteria for causality and it is summarized in the experience of the International Agency for Research on Cancer (IARC) Monographs programme whose systematic review methodology has been recently updated (Samet et al. 2020). In addition, the Integrated Science Assessment of the US Environmental Protection Agency (EPA) (US EPA, 2016; 2019) represents a framework with specific application to air pollutants.

The Panel felt that some important considerations (e.g., number of studies, strength of the association), as well as the methodology employed in the IARC evaluations and US EPA Integrated Science Assessments were not well considered in the OHAT approach. The main reason for this was that the main purpose in OHAT is to assess confidence in the quality of the body of evidence, rather than to assess confidence in the presence of an association, i.e., OHAT focus more on the quality of the studies rather than on their results. In addition, the OHAT approach was heavily geared towards the studies entering a meta-analysis rather than considering all the available evidence. The limitation that meta-analysis should include only studies that are sufficiently compatible to be pooled into an aggregate estimate has been noted (Savitz and Forastiere, 2021). Therefore, the Panel thought it was prudent to accompany the modified OHAT assessment with a broader approach to assess the level of confidence in the presence of an association. To this end, the Panel developed a narrative assessment for each health outcome. In the narrative assessment, all studies identified in the systematic search were considered whether included in the meta-analyses or not. The narrative assessment also included studies with indirect traffic measures such as distance and traffic density.

The narrative assessment included the following aspects: evaluation of the number, size, and location of the evidence base; study design, study population and representativeness, the strength (magnitude) and nature (direction) of the association, quality of the studies (e.g., confounding, selection bias, exposure assessment, outcome assessment, missing data and selective reporting); consistency of the findings, e.g., across locations, age groups, time periods, study designs, and different pollutants and indirect traffic measures, traffic specificity and adjustment for noise for some outcomes); monotonic exposure-response function, and other considerations. The results of the meta-analyses, as well as the findings from studies not in the meta-analysis, were important for the evaluation, as a larger relative risk (with narrow confidence intervals) was more likely to indicate an association with TRAP than was a smaller and uncertain effect estimate. Associations that were replicated in several studies of the same design, across different populations, or across several pollutants, or that used different epidemiological approaches or under different circumstances of exposure were more likely to represent a true association than isolated observations from single, small studies. The presence of a monotonic exposure-response function was considered a strong indication of an association. In this way, the narrative assessment took into consideration all the available evidence from both the metanalytic results and the results of single studies without a meta-analysis and assessed the level of confidence in the evidence that TRAP is associated with the selected health outcome.

The narrative assessment of the level of confidence in the presence of an association between TRAP and a specific outcome were summarized as high (large number of studies, confounding, other biases and chance can be reasonably excluded, consistent associations across multiple populations and pollutants), moderate (moderate/large number of studies, confounding, other biases and chance cannot be reasonably excluded, moderate consistency of associations across populations and pollutants), low (small number of studies, confounding, other biases and chance are likely, inconsistency of associations across populations and pollutants) or very low (small number of studies, confounding, other biases and chance very likely and large inconsistencies of associations across populations and pollutants). These considerations were not applied automatically with set criteria for the issues considered.

Table 9 presents a comparison of main similarities and differences between the narrative assessment and the modified OHAT assessment. Both assessments were considered complementary, reflecting the complex issues in determining the level of confidence.

Table 9. Comparison of main similarities and differences between the narrative assessment and the modified OHAT assessment.

	Narrative assessment	**Modified OHAT assessment**
Main purpose	to assess confidence in the presence of an association	to assess confidence in the quality of the body of evidence
Inclusion of studies	All studies - both the meta-analytic results and results of studies that were not included in meta-analysis	All studies, though heavily geared towards the studies entering a meta-analysis
Number, location, and size of the evidence base	Yes	Partial
Study design	Yes	Yes
Study population (generalizability)	Yes	No
Strength and nature of the association	Yes	No[1]
Risk of bias	Yes	Yes
confounding	Yes	Yes
selection bias	Yes	Yes
exposure assessment	Yes	Yes
outcome assessment	Yes	Yes
missing data	Yes	Yes
selective reporting	Yes	Yes
Consistency of the findings (e.g., across locations, time periods, study designs, and different pollutants and indirect traffic measures)	Yes	Partial
Unexplained inconsistency	Yes	Yes
Imprecision (chance)	Yes	Yes
Publication bias	No	Yes
Exposure-response	Yes	Yes
Residual confounding	Yes	Yes

[1]The OHAT has an upgrading factor for "large magnitude of effect" that applies only if the effect size is large or very large (i.e., large RR > 2 or very large RR > 5) because residual confounding is then less likely. However, the Panel consider a "large" effect to be both ambiguous to define and unlikely to occur. Thus, the Panel has decided not to consider this specific upgrading factor.

5.3.5 Overall confidence assessment

The confidence assessment of the narrative assessment and the assessment based on the modified OHAT approach was combined in an overall assessment between TRAP and the selected health outcomes. 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 listed both (e.g., "moderate to high", since the Panel considered both assessments complementary, reflecting the complex issues in determining the level of confidence. Detailed descriptors of the level of the evidence for an association are listed in Table 10.

Table 10. Overall confidence assessment: descriptions of the level of confidence in the evidence for an association[1].

High	Evidence is sufficient to conclude that the strength of the evidence for an association is high, that is, the exposure has been shown to be associated with health effects in studies in which chance, confounding, and other biases could be ruled out with reasonable confidence. The determination is based on multiple high-quality studies conducted in different populations and geographical areas with consistent results for multiple exposure indicators. High confidence in the association between exposure and the outcome.
Moderate	Evidence is sufficient to conclude that an association is likely to exist, that is, the exposure has been shown to be associated with health effects in studies where results are not explained by chance, confounding, and other biases, but uncertainties remain in the evidence overall. The determination is based on some high-quality studies in different populations and geographical areas, but the results are not entirely consistent across areas and for multiple exposure indicators. Moderate confidence in the association between exposure and the outcome.
Low	Evidence is suggestive but limited, and chance, confounding, and other biases cannot be ruled out. Generally, the body of evidence is relatively small, with few high- quality studies available and at least one high-quality epidemiologic study shows an association with a given health outcome and/or when the body of evidence is relatively large but the evidence from studies of varying quality and across multiple exposure indicators is generally supportive but not entirely consistent. Low confidence in the association between exposure and the outcome.
Very low	Evidence is inadequate to determine if an association exists with the relevant exposures. The available studies are of insufficient quantity, quality, consistency, or statistical power to permit a conclusion regarding the presence or absence of an association. Very low confidence in the association between exposure and the outcome.

[1]The overall assessment of the association of each health outcome with long-term exposure to TRAP is a combination of the narrative assessment and the modified OHAT assessment. The descriptors are modified from US EPA 2015 and OHAT 2019a.

5.3.6 References

Agency for Healthcare Research and Quality (AHRQ). 2013. Grading the Strength of a Body of Evidence When Assessing Health Care Interventions for the Effective Health Care Program of the Agency for Healthcare Research and Quality: An Update. AHRQ Publication No. 13(14)-EHC130-EF. Rockville MD: AHQR.

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; doi:10.1186/s13643-018-0915-2

Chen J, Hoek G. 2020. Long-term exposure to PM and all-cause and cause-specific mortality: A systematic review and meta-analysis. Env Int; doi:10.1016/j.envint.2020.105974.

Dekkers OM, Vandenbroucke JP, Cevallos M, Renehan AG, Altman DG, Egger M. 2019. COSMOS-E: Guidance on conducting systematic reviews and meta-analyses of observational studies of etiology. PLoS Med; doi:10.1371/journal.pmed.1002742.

Guyatt GH, Oxman AD, Kunz R, et al. 2011. GRADE guidelines 6. Rating the quality of evidence--imprecision. J Clin Epidemiol; doi:10.1136/bmj.39489.470347.AD.

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: Health Effects Institute.

Huangfu P, Atkinson R. 2020. Long-term exposure to NO2 and O3 and all-cause and respiratory mortality: A systematic review and meta-analysis. Environ Int; doi:10.1016/j.envint.2020.105998.

Ioannidis JPA, Tan YJ, Blum MR. 2019. Limitations and misinterpretations of E-Values for sensitivity analyses of observational studies. Ann Intern Med; doi:10.7326/M18-2159

Lawlor DA, Tilling K, Davey Smith G. 2016. Triangulation in aetiological epidemiology. Int J Epidemiol 2016; doi:10.1093/ije/dyw314.

Lin L, Chu H. 2018.Quantifying publication bias in meta-analysis. Biometrics; doi:10.1111/biom.12817.

Liu Y, Sun J, Gou Y, Sun X, Li X, Yuan Z, et al. 2018. A multicity analysis of the short-term effects of air pollution on the chronic obstructive pulmonary disease hospital admissions in Shandong, China. Int J Environ Res Public Health; doi:10.3390/ijerph15040774.

Maldonado G, Greenland S. Estimating causal effects. 2002. Int J Epidemiol; doi:10.1093/ije/31.2.422.

Morgan RL, Thayer KA, Santesso N, Holloway AC, Blain R, Eftim SE, et al. 2019. GRADE Working Group. A risk of bias instrument for non-randomized studies of exposures: A users' guide to its application in the context of GRADE. Environ Int; doi:10.1016/j.envint.2018.11.004

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; doi:10.1136/bmjebm-2018-110891

Office of Health Assessment and Translation (OHAT). 2019a. Handbook for Conducting a Literature– Based Health Assessment Using OHAT Approach for Systematic Review and Evidence Integration. Division of the National Toxicology Program. National Institute of Environmental Health Sciences. Available:

https://ntp.niehs.nih.gov/ntp/ohat/pubs/handbookmarch2019_508.pdf [accessed October 12, 2021].

Office of Health Assessment and Translation (OHAT). 2019b. NTP Monograph on Systematic Review of Traffic-Related Air pollution and hypertensive disorders of pregnancy. Research Triangle Park, NC:National Toxicology Program, Public Health Services, U.S. Dept. of Health and Human Services. Available: *https://ntp.niehs.nih.gov/ntp/ohat/trap/mgraph/trap_final_508.pdf* [accessed October 12, 2021].

Pearce N, Vandenbroucke JP, Lawlor DA. 2019. Causal inference in environmental epidemiology: Old and new approaches. Epidemiology; doi:10.1097/EDE.0000000000000987

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; doi:10.1289/ehp.1307972

Rothman KJ, Greenland S. 2018. Planning study size based on precision rather than power. Epidemiology; doi:10.1097/EDE.0000000000000876.

Rücker G, Schwarzer G, Carpenter JR, et al. 2008. Undue reliance on I^2 in assessing heterogeneity may mislead. BMC Med Res Methodol; doi:10.1186/1471-2288-8-79

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; doi:10.1093/jnci/djz169.

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: Assess the impact of specific sources of potential bias. Am J Epidemiol; doi:10.1093/aje/kwz131.

Savitz DA, Forastiere F. 2021. Do pooled estimates from meta-analyses of observational epidemiology studies contribute to causal inference? Occup Environ Med; doi:10.1136/oemed-2021-107702.

Schünemann H, Brożek J, Guyatt G, Oxman AD, eds. 2013. GRADE Handbook. Available: *https://gdt.gradepro.org/app/handbook/handbook.html* [accessed 21 May 2021].

Shi L, Lin L. 2019. The trim-and-fill method for publication bias: Practical guidelines and recommendations based on a large database of meta-analyses. Medicine; doi:10.1097/MD.0000000000015987.

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; doi:10.1289/EHP6980

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.

U.S. EPA (U.S. Environmental Protection Agency). 2016. Integrated Science Assessment for Oxides of Nitrogen–Health Criteria. EPA/600/R-15/068. Research Triangle Park, NC:U.S. EPA.

U.S. EPA (U.S. Environmental Protection Agency). 2019. Integrated Science Assessment for Particulate Matter. EPA/600/R-19/188. Research Triangle Park, NC:U.S. EPA.

VanderWeele TJ, Ding P. 2017. Sensitivity Analysis in Observational Research: Introducing the E-Value. Ann Intern Med; doi:10.7326/M16-2607.

WHO (World Health Organization). 2020. Risk of Bias Assessment Instrument for Systematic Reviews Informing WHO Global Air Quality Guidelines. Bonn:WHO Regional Office for Europe. Available: *https://www.euro.who.int/en/health-topics/environment-and-health/air-quality/publications/2020/risk-of-bias-assessment-instrument-for-systematic-reviews-informing-who-global-air-quality-guidelines-2020* [accessed October 12, 2021].

WHO (World Health Organization. 2018. Environmental Noise Guidelines for the European Region. Bonn:WHO Regional Office for Europe.

Woodward M. 2013. Epidemiology: Study Design and Data Analysis. Third Edition. Boca Raton, FL:CRC Press, a Taylor and Francis group.

Zhang Y, Coello PA, Guyatt GH, Yepes-Nuñez JJ, Akl EA, Hazlewood G, Pardo-Hernandez H, Etxeandia-Ikobaltzeta I, Qaseem A, Williams JW Jr, Tugwell P, Flottorp S, Chang Y, Zhang Y, Mustafa RA, Rojas MX, Xie F, Schünemann HJ.2019. GRADE guidelines: 20. Assessing the certainty of evidence in the importance of outcomes or values and preferences-inconsistency, imprecision, and other domains. J Clin Epidemiol; doi:10.1016/j.jclinepi.2018.05.011.