RISK OF BIAS ASSESSMENT INSTRUMENT FOR SYSTEMATIC REVIEWS INFORMING WHO GLOBAL AIR QUALITY GUIDELINES

By: the WHO Global Air Quality Guidelines
Working Group on Risk of Bias Assessment

Abstract

Air pollution represents the single largest environmental threat to health globally. WHO is developing new global air quality guidelines, based on the most recent scientific evidence, synthetized in accordance to predefined systematic review methods. This publication includes a description of the instrument devised to assess the risk of bias in the individual studies included in the systematic reviews of adverse health effects informing the guidelines. Further, general instructions and specific subject matter advice to guide the assessments are provided.

Contents

Acknowledgements

Abbreviations

AQGs	air quality guidelines
GRADE	Grading of Recommendations Assessment, Development and Evaluation (framework)
PECOS	Population, Exposure, Comparator, Outcome(s), and Study Design
PM	particulate matter
$PM_{2.5}$	particulate matter with a diameter of 2.5 μm or less
ROBINS-I	Risk of Bias In Non-randomized Studies - of Interventions
SES	socioeconomic status

Background

With the release of the Risk of Bias In Non-randomized Studies - of Interventions (ROBINS-I) *(1)* and the subsequent publication of an equivalent instrument to evaluate ROBINS-I of exposures *(2)*, research is progressing to assist systematic review authors and guideline developers. For guideline panels and decision-makers, researchers are defining methods for integration of these instruments into established evidence assessment frameworks, such as the Grading of Recommendations Assessment, Development and Evaluation (GRADE) approach *(3)*. However, interest exists to further develop the current instrument for assessment of risk of bias in observational air pollution epidemiology studies. Likewise, recent guidance on systematic review of observational studies of aetiology recommends the development of tailored-made instruments for each review, based on general principles *(4)*.

Assessment of potential bias lies at the core of all epidemiology when trying to understand the relationship between an exposure and an outcome. In addition, it is a basic component of any well-performed systematic review of evidence, as outlined by the *WHO handbook for guideline development, 2nd edition (5)*. When conducting systematic reviews to inform the WHO global air quality guidelines (AQGs), risk of bias assessment was therefore necessary for all eligible studies to understand the certainty in that relationship.

The current instrument was iteratively developed by members of the Guideline Development Group for the update of WHO global AQGs with expertise in air pollution epidemiology and experts in developing and applying risk of bias instruments in observational studies, under the oversight of WHO staff. The instrument was designed specifically to evaluate risk of bias within eligible air pollution studies included in systematic reviews, commissioned by WHO, of studies on short- and long-term exposure to air pollutants (i.e. particulate matter (PM) smaller than 2.5 and 10 µm in diameter, nitrogen dioxide, sulphur dioxide, carbon monoxide and ozone), to examine their effect on critical patient-important outcomes, such as natural cause and cause-specific mortality (e.g. cardiovascular mortality).

Studies of short- and long-term exposure to air pollution use different study designs and make use of different contrasts of exposure: day-to-day fluctuation for the short-term studies and spatial variation of average concentrations for the long-term exposure studies. Designs included in the WHO-commissioned systematic reviews were time series, case-crossover and panel studies (short-term) and cohort and case-control studies (long-term).

In developing this instrument, each topic of interest (exposures and outcomes) was examined based on the research questions (framed as PECOS; Population, Exposure, Comparator, Outcome(s), and Study Design) to inform the confounders and analysis methods that should be taken into account when determining the potential for bias within a primary study. In a well-formed research question (such as a PECOS), a distinction is made between an exposure and comparator group. This is required for decision-making, as the exposure and the comparator distinguish between the options available. For studies of exposure, the PECOS may take one of five variations *(6)*. For WHO global AQGs, the interest lies in the comparison between different levels of exposure (e.g. a lower level compared to a higher level of exposure to $PM_{2.5}$).

To inform the development process, WHO staff conducted a review of existing tools at inception. In addition, members of the Working Group on Risk of Bias Assessment and the Systematic Review Team pilot tested the instrument for accuracy. It is important to note, however, that the risk of bias instrument below is not a checklist, but a list of guided topics organized by domains and subdomains to consider when judging the potential for bias. The instrument should be used by assessors with detailed subject matter knowledge of air pollution epidemiology and training in risk of bias instrument application.

Risk of bias assessment instrument

Risk of bias instrument	Topic:			Reviewer ID:		
		Study ID:				
		Date:				

For each PECOS		Long-term studies	Short-term studies			Notes
	Critical potential confounders					
	Other potential confounders					

Domain	Subdomain	Low-risk (ideal study) criteria	Moderate-risk criteria	High-risk criteria	Overall judgement for a domain: Low/ Moderate/ High	Rationale/ Notes (quotes from the study to justify the judgement)
1. Confounding	Were all confounders considered adjusted for in the analysis?	All critical and other/additional potential confounders adjusted for or 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).	All critical potential confounders but not all other/ additional potential confounders adjusted for without 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).	Not all critical potential confounders adjusted for without support (e.g. exploratory analysis) of minimal risk due to residual confounding.		
	Validity of measuring of confounding factors	Confounders measured with documented valid methods.	Not all critical potential confounders were measured with documented valid methods; however, there is evidence that this does not lead to severe confounding.	Any critical or other/additional potential confounder not validly assessed and evidence of residual confounding.		
	Control in analysis (Did the authors use an appropriate analysis method or study design that controlled for confounding domains?)	Authors used appropriate analysis methods or study designs that controlled for confounding domains.	Authors used inappropriate methods or designs when adjusting for critical potential confounders; however, there is evidence that this does not lead to severe confounding.	Authors used inappropriate methods or study designs when adjusting for critical and other/additional potential confounders.		
				Overall		

contd.

Domain	Subdomain	Low-risk (ideal study) criteria	Moderate-risk criteria	High-risk criteria	Overall judgement for a domain: Low/ Moderate/ High	Rationale/ Notes (quotes from the study to justify the judgement)
2. Selection bias	Selection of participants into the study (includes 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 (rationale required).	Participants in all exposure levels did not have equal opportunity to be in the study, to the extent that effect estimates were seriously biased.		
				Overall		
3. Exposure assessment	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.		
	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.		
	Change in exposure status (for long-term studies only)	Spatial exposure contrasts did not change throughout the study or time varying exposure was used to account for changes.	Spatial exposure contrasts did change throughout the study and were not accounted for **but** effect estimates were not seriously biased.	Spatial exposure contrasts did change throughout the study and were not accounted for, **and** effect estimates were seriously biased and were different in cases and non-cases.		
	Exposure contrast	Exposure contrast was large compared to the precision of exposure assessment (between-subject variance larger than within-subject variance).	Exposure contrast was small relative to the within-subject variance but not to the extent that the study is uninformative.	Exposure contrast was so small relative to the within-subject variance that the study is uninformative.		
				Overall		

contd.

Domain	Subdomain	Low-risk (ideal study) criteria	Moderate-risk criteria	High-risk criteria	Overall judgement for a domain: Low/ Moderate/ High	Rationale/ Notes (quotes from the study to justify the judgement)
4. Outcome measurement	Blinding of outcome measurement	Outcome measurements 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.		
	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.		
	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.		
				Overall		
5. Missing data	Missing data of outcome measures	No missing outcome data or missing data infrequent (<10%) or missing data related to outcome or exposure data imputed using appropriate methods.	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.		
	Missing data of exposures	No missing exposure data or missing data infrequent (<10%) or missing data related to exposure or outcome data imputed using appropriate methods.	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.		
				Overall		

contd.

Domain	Subdomain	Low-risk (ideal study) criteria	Moderate-risk criteria	High-risk criteria	Overall judgement for a domain: Low/ Moderate/ High	Rationale/ Notes (quotes from the study to justify the judgement)
6. Selective reporting	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.		
				Overall		

General instructions to begin the risk of bias assessment

The following fields should be populated in the risk of bias instrument:

- topic of the review (refer to the specific PECOS from the systematic review)
- date when the risk of bias assessment is conducted
- reviewer ID (coordinate across the systematic review team)
- study ID (coordinate across the systematic review team and studies)
- relevant critical and other/additional potential confounders.

Conduct risk of bias assessment at the outcome level

The risk of bias assessment is conducted at the outcome level; therefore, should a primary study report on two relevant outcomes, risk of bias must be evaluated twice. This is because the risk of bias may be different depending on the outcome. For example, if the same study reports on both all-cause mortality and cardiovascular mortality, separate risk of bias evaluations must be conducted. The risk of bias assessment of all-cause mortality may differ from that of cardiovascular mortality depending on the certainty of the ability to differentiate specific causes of mortality. It is important to clearly state when there is uncertainty around the identification or diagnosis of the outcome of interest.

Provide a rationale for all judgements

As mentioned previously, this instrument is not a checklist because different judgements can be made as to the potential for bias introduced by each subdomain. As such, for the transparency of the overall risk of bias for the systematic review and subsequent guideline, rationales must be provided for each risk of bias judgement made. The rationale should address why the specific judgement ("low", "moderate" or "high" risk of bias) was selected. The rationale may be in a narrative form from the rater or use a quote from the primary study under review.

Identify all critical and other potential confounders prior to assessing risk of bias

A confounder is a factor associated with the exposure that influences the disease outcome *(7)*. Before raters attribute any difference in the outcome between the exposure and comparison to the exposure of interest, exploration of potential factors along this causal pathway (i.e. confounders) is required. Confounders have a relationship with the outcome and exposure, and their presence (unless adequately adjusted for) can lead to a mis-specified (either over- or underestimate) association between the outcome and exposure.

Typically, subject matter expertise is needed when identifying critical and other potential confounders. Critical potential confounders may be thought of as strong determinants of the health outcome at hand (e.g., smoking and lung cancer), while other potential confounders would be less strong/consistent determinants (e.g., diet and chronic obstructive pulmonary disease). However, some other potential confounders may be considered critical in that they are powerful known risk factors for the outcome, and lack of adjustment for them (by analysis or restriction) is a major source of uncertainty. Critical and other potential confounders that should be adjusted for in each primary study should be identified in the initial stages of the systematic review. However, there may be situations during the conduct of the systematic review where the reviewers realize that either: (i) the confounders

are not relevant or (ii) a confounder that was not identified should indeed be captured in the study. The Working Group identified confounders for the specific reviews informing AQGs in Table 1. This is a task specific to each body of evidence at hand.

Table 1. Critical and other potential confounders to assess confounding in systematic reviews informing WHO global AQGs

Critical potential confounders	Other potential confounders
Long-term studies: cohort, case-control	
- Age - Sex - Individual- or area-level socioeconomic status (SES) (at least one of the following): – educational level (individual) – employment status (occupational class) (individual) – individual income – mean income (area-level) – deprivation index (area-level) or other SES indices - Body mass index	- Year of enrolment - Ethnicity - Diet - Physical activity - Marital status - Smoking
Short-term studies: time series, case crossover,[a] panel	
- Temperature - Seasonality - Day of the week - Long-term trends	- Holidays (not vacations) - Influenza epidemics

[a] The case-crossover approach may control for seasonality and long-term trends by design.

Specific guidance to complete the risk of bias assessment

For each domain of risk of bias, related subdomains and guidance are provided to assist raters in making a judgement about whether the study presents "low", "moderate" or "high" risk of bias. Since raters will make a judgement as to the risk of bias for each subdomain, it is important that a rationale is provided in the risk of bias instrument.

Confounding

In addition to the guidance on confounding provided above, it is important to consider some issues on controlling for long-term trends.

While the PECOS may be asking for the relationship between the exposure of interest and short-term variation in the outcome, long-term trends may dominate the potential confounders. It is therefore essential to demonstrate a separation between short- and long-term trends to have certainty in the assessment of the relationship between the exposure and outcome. Regression models may be appropriate to demonstrate that long-term trends are controlled for in the analysis. Some options available include: (i) a time stratified model (simple indicator variables); (ii) periodic functions; or (iii) flexible spline functions. The use of one of these models to adjust for potential long-term confounders allows for the isolation of the exposure of interest *(8)*.

It is also necessary to think in which direction the long-term trend would potentially bias the effect (i.e. toward or away from the null).

Selection bias

Selection bias occurs when some eligible participants are excluded in a way that leads to the association between exposure and outcome differing from the association that would have been observed in the PECOS.

Exposure assessment

The variable of interest in the exposure measurement domain is the personal exposure to outdoor pollutants as the outdoor concentration near the homes of studied populations. Most people spend over 70% of their time at home. A high percentage of this time is spent indoors, so a legitimate question is how well spatial and temporal variations in outdoor concentrations are correlated with spatial and temporal variations in indoor and personal exposure to outdoor pollutants. A sizeable literature shows these correlations to be moderately to very high. As personal or indoor monitoring is impossible in the typically large study populations in air pollution epidemiology, the focus is therefore on assessment of the risk of bias associated with the various ways in which outdoor concentrations have been assessed in the studies included in the systematic reviews. However, raters should also be aware – for the case of $PM_{2.5}$ – of the systematic reviews of residential outdoor-personal and ambient-personal correlations by Avery et al., 2010 *(9,10)*. Again here the basic idea is to make a distinction between studies at high risk of bias because the assessment of exposure in the study at hand was inadequate, and studies at low risk of bias because the exposure assessment was performed in an adequate way.

Studies of the effects of short-term, day-to-day variations in air pollution on the day-to-day variations in mortality and hospital admissions typically rely on routine measurements of air pollutants at a limited number of urban or regional background sites. Raters should consider the following questions.

- Is the quality assurance of the monitoring instruments documented? For example, the study uses data from national agency networks with established quality assurance procedures or a comparable quality assurance

procedure (low risk of bias); residence or employment is in a defined geographical area where exposure can be assumed (high risk of bias).

- Are the monitoring sites representing the day-to-day variations of outdoor conditions of the study populations of interest? For example, data are restricted to urban or regional background stations (low risk of bias) or data are influenced significantly by a particular source of air pollution, such as a busy traffic street or an industrial site (high risk of bias).

- How are criteria for completeness reported? For example, >18 hours/day, >75% days/month, etc. have a low risk of bias; ≤ 18 hrs/day, ≤ 75% days/month, etc. have a high risk of bias.

Day-to-day variations in outdoor air pollution concentrations are typically large (> one order of magnitude) and largely driven by meteorology; thus, these day-to-day variations are not very different between different pollutants of interest in the AQG context.

Studies of the effects of long-term, spatial variations in air pollution on the spatial variation in mortality and morbidity outcomes rely on various measurements and modelling efforts. Spatial variations in long-term air pollution concentrations are driven by spatial variations in source strength, road networks, topography, etc. Interpolation between data from routine monitoring stations may be adequate in some studies able to use dense monitoring networks. In most cases, some combination of monitoring data, dispersion modelling, land use regression modelling and satellite observations is being used to estimate the relevant spatial variations. Scales vary from tens of metres for the finest scale land-use regression and dispersion models to 10 x 10 km for some satellite products. Issues for the raters to consider are:

- the ability of the exposure models used in the studies to adequately predict the exposure; this can be concluded if the model is adequately evaluated against measurements (low risk of bias) or is not evaluated against measurements (high risk of bias); and

- the temporal stability over time scales relevant for the long-term studies of interest; e.g., if the exposure contrast is generated for a specific year, it is representative for other years of the epidemiological study and outcome of interest (low risk of bias) or if it is unrepresentative (high risk of bias).

Typically, details such as these are found in separate technical papers; the raters should ascertain that they are referenced.

Outcome measurement

Questions within this domain are to determine whether bias has been introduced by the misclassification or introduction of error in the measurement of the outcome. This may differ depending on the outcome of interest and should be examined separately for each outcome.

Missing data

This domain aims to capture the potential for bias introduced when not all data are available for the analysis and the missing data are related to both the exposure and the outcome. This attrition could be due to loss to follow up, missing appointments or exclusion by the study coordinators. The impact from missing data should be informed by considering best- and worst-case situations (i.e. all missing persons are in the exposed group instead of the comparison group vs no missing persons are in the exposed group; instead all are in the comparison group) to determine the impact on the overall effect of the exposure. Similarly, it is important to consider the potential impact when the missing data are different among the exposed and the comparator groups.

Selective reporting

Raters should consider if there are any reasons to suspect that the results reported are different or selected from the results intended to be measured. This may be identified by referring to the protocol of the study to determine if the primary outcomes are consistently reported in all documents. This may also be identified if only a subgroup

from a larger cohort is reported. When a previously published protocol is not available, sometimes the methods section may be the best indicator of the a priori research plan.

Overall judgement for a domain

To come to an overall judgement for a domain, the following procedure applies: if any of the subdomains has a rating of high risk of bias, the whole domain is rated as high risk of bias; if all the subdomains have a rating of low risk of bias, the whole domain is rated as low risk of bias; when at least one subdomain has a rating of moderate risk of bias and none of the other subdomains is at high risk of bias, the whole domain is rated as moderate risk of bias.

Judgements per risk of bias domain across studies

To avoid carrying forward the ratings from one domain to the others, the Working Group considered that an overall judgement of bias at the study level was not appropriate. Instead, subgroup analyses should performed per risk of bias domain across studies, grouping studies at higher risk of bias for that domain and studies at lower risk of bias for that domain *(11)*. This would then give a better idea which bias has an effect on the effect size, in what direction (over- or underestimation of the true effect) the bias lays and how much it matters in the presentation of the effect estimate.

Subgroup analyses may have little power of detecting differences when there are only a limited number of studies (=cases) involved. In addition, these are all ecological analyses, meaning that comparisons are made at the group level, whereas comparisons are of interest at the participant level.

Instead of subgroup analyses, the raters can also conduct a meta-regression, but this is basically the same procedure as described above. The advantage is that more variables than risk of bias can be brought into a regression equation. However, it is unusual to look at more than risk of bias in subgroup analyses.

References

1. Sterne JA, Hernan MA, Reeves BC, Savovic J, Berkman ND, Viswanathan M, et al. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. BMJ. 2016;355:i4919. doi:10.1136/bmj.i4919.

2. Morgan RL, Thayer KA, Santesso N, Holloway AC, Blain R, Eftim SE, et al. Evaluation of the risk of bias in non-randomized studies of interventions (ROBINS-I) and the 'target experiment'concept in studies of exposures: Rationale and preliminary instrument development. Environ Int. 2018;120:382–7. doi:10.1016/j.envint.2018.08.018.

3. Schünemann HJ, Cuello C, Akl EA, Mustafa RA, Meerpohl JJ, Thayer K, et al. GRADE guidelines: 18. How ROBINS-I and other tools to assess risk of bias in nonrandomized studies should be used to rate the certainty of a body of evidence. J Clin Epidemiol. 2019;111:105–14. doi:10.1016/j.jclinepi.2018.01.012.

4. Dekkers OM, Vandenbroucke JP, Cevallos M, Renehan AG, Altman DG, Egger M. COSMOS-E: Guidance on conducting systematic reviews and meta-analyses of observational studies of etiology. PLoS Med. 2019;16(2):e1002742. doi:10.1371/journal.pmed.1002742.

5. WHO handbook for guideline development, 2nd edition. Geneva: World Health Organization; 2014 (http://apps.who.int/medicinedocs/en/m/abstract/Js22083en/, accessed 18 November 2019).

6. Morgan RL, Whaley P, Thayer K, Schünemann H. Identifying the PECO: A framework for formulating good questions to explore the association of environmental and other exposures with health outcomes. Environ Int. 2018;121(Pt 1):1027–31. doi:10.1016/j.envint.2018.07.015.

7. Kirkwood BR, Sterne JA. Essential Medical Statistics. Chichester: John Wiley & Sons; 2010.

8. Bhaskaran K, Gasparrini A, Hajat S, Smeeth L, Armstrong B. Time series regression studies in environmental epidemiology. Int J Epidemiol. 2013;42(4):1187–95. doi:0.1093/ije/dyt092.

9. Avery CL, Mills KT, Williams R, McGraw KA, Poole C, Smith RL, et al. Estimating error in using ambient PM2.5 concentrations as proxies for personal exposures: a review. Epidemiology. 2010: 21(2): 215–23. doi:10.1097/EDE.0b013e3181cb41f7.

10. Avery CL, Mills KT, Williams R, McGraw KA, Poole C, Smith RL, et al. Estimating error in using residential outdoor PM2.5 concentrations as proxies for personal exposures: a meta-analysis. Environ Health Perspect. 2010;118(5):673–8. doi:10.1289/ehp.0901158.

11. Morgan RL, Thayer KA, Santesso N, Holloway AC, Blain R, Eftim SE, et al. A risk of bias instrument for non-randomized studies of exposures: A users' guide to its application in the context of GRADE. Environ Int. 2019;122:168–84. doi:10.1016/j.envint.2018.11.004.