Approach to assessing the certainty of evidence from systematic reviews informing WHO global air quality guidelines

By: the WHO Global Air Quality Guidelines Working Group on Certainty of Evidence Assessment

Acknowledgements

This supplementary material consists of an approach to assessing the certainty of evidence from systematic reviews of epidemiologic studies of air quality and health, based on the Grading of Recommendations Assessment, Development and Evaluation (GRADE) framework.

The approach was developed by external methodologist Jos Verbeek (Cochrane Work), with inputs from the WHO Global Air Quality Guidelines Working Group on Certainty of Evidence Assessment, convened by the WHO European Centre for Environment and Health (WHO Regional Office for Europe) in the context of the forthcoming WHO global air quality guidelines. The Working Group was composed of the Guideline Development Group members: Aaron Cohen (Health Effects Institute), Bert Brunekreef (Utrecht University), Francesco Forastiere (King's College London), Nino Künzli (Swiss Tropical and Public Health Institute), and external methodologist: Rebecca Morgan (McMaster University); and, from the staff of the WHO Regional Office for Europe: Román Pérez-Velasco, Hanna Yang and Dorota Jarosińska. Additional comments were provided at different stages by external methodologist Eva Rehfuess (Cochrane Public Health Europe) and GDG members Michal Krzyzanowski (King's College London), and Jonathan Samet (Colorado School of Public Health).

The WHO Regional Office for Europe acknowledges funding and in-kind contributions from the European Commission (Directorate-General for Environment); the German Federal Ministry for the Environment, Nature Conservation and Nuclear Safety; the German Federal Ministry of Health; the Government of the Republic of Korea; the Swiss Federal Office for the Environment; and the United States Environmental Protection Agency.

Background

The Grading of Recommendations Assessment, Development and Evaluation (GRADE) has been developed to standardize the approach to judging the certainty of the effects of interventions *(1)*. As such, the approach is currently the basis for evidence review in support of WHO Guidelines *(2)*.

The main value of the system is that the comparability of the judgements increases when all assessors consider the same arguments underpinning their certainty in a similar manner. That is how the factors for downgrading and upgrading the certainty have been developed: to guide expert judgement. Behind each down- and upgrading factor in GRADE, there is a rationale for its importance and guidance for elaborating good reasons for downgrading or not downgrading. These ideas are well explained in the *GRADE Handbook (1)*. Most of the reasoning in this framework can be equally well used for observational studies of exposure as for randomized studies of interventions *(3)*. However, at some points there is a need for elaboration or clarification on how to use the GRADE criteria for observational studies of exposure.

Although different groups have adapted the approach for environmental exposures in recent years, no consensus has emerged among experts yet. Unlike some previous efforts, the aim of this work is not assessing the strength of evidence for causal inference by considering all the relevant streams of research *(4)*, but to rate how certain one is that the 'true' estimate of the epidemiological association between an air pollutant and an adverse health effect lies within a particular range *(5)*. Consistent with the standard GRADE framework, the certainty of the effect estimate is graded as high, moderate, low or very low. The ratings are subsequently used to select and underpin concentration–response functions in the process of deriving guideline exposure levels.

The current approach was designed specifically to assess the certainty of the evidence from the systematic reviews commissioned by WHO to inform the update of global air quality guidelines (AQGs). Its development benefitted from previous experiences in applying GRADE in the field of occupational and environmental health, as well as specific expertise in air pollution epidemiology. The approach was extensively discussed in two Guideline Development Group meetings, pilot tested by the members of the Systematic Review Team and improved iteratively according to the feedback received.

The Working Group accepted to start the rating of the certainty of the evidence for observational studies at moderate certainty evidence and not at high certainty, because of the risk of unmeasured confounding in observational studies. The certainty of the evidence from this level can then be downgraded or upgraded, based on the criteria per GRADE domain. The GRADE domains and the criteria considered when judging the certainty of the evidence are elaborated below.

Reasons for downgrading

Limitations in studies: downgrade one or two levels

For risk of bias in studies, there should be serious concern about bias in the studies that have the most weight in the meta-analysis to rate down the certainty of the total body of evidence with one level. If there are very serious concerns, the certainty can be downgraded with two levels.

This is a judgement and there are no clear pre-set cut-off points *(6)*. A judgement is based on the number of studies and the impact they have in the meta-analysis, as well as the seriousness of the risk of bias in these studies. One small study with very serious risk of bias but hardly an influence on the meta-analysis should not be a reason to downgrade, but two big studies with a considerable weight in the meta-analysis should.

If the sensitivity analysis for risk of bias shows a considerable impact on the effect size, the conclusions could be based on the studies at low risk of bias only. In that case, there is no reason to downgrade because the body of evidence on which the conclusions are based is considered to be at low risk of bias only.

Indirectness: downgrade one or two levels

The assessors should consider the extent to which the Population, Exposure, Comparator, Outcome(s), Study Design (PECOS) of the studies in the meta-analysis reflects the original PECOS question formulated at the beginning of the systematic review process *(7)*.

If there are considerable differences between the elements of the PECOS in the body of evidence compared to the original question, then the certainty of the body of evidence should be rated down with one level. This would, for example, be the case if the evidence consists of studies of occupational exposure instead of exposure in the general population.

Inconsistency: downgrade one or two levels

Inconsistency among studies means that there is a considerable difference in effect size between studies. For example, if there are studies in the body of evidence that show a harmful effect and also studies that show a preventive effect, this indicates serious inconsistency or heterogeneity.

Usually there is more heterogeneity in observational than in experimental studies, because more factors can influence the effect size. Therefore, it is important to try to explain the heterogeneity. The first step should be to consider the factors that are listed for subgroup analyses in the protocol, as those that are most likely to be moderators of effect sizes. Another source of heterogeneity can be variation in risk of bias. This may explain part of the heterogeneity, and evaluation of only studies at low risk of bias should then decrease the heterogeneity. The difference in effect sizes between the subgroups should be tested for statistical significance. A rule of thumb to be used is to check if the confidence intervals of the subgroup pooled effect sizes do not overlap.

Ideally, a meta-regression should be conducted including all moderators of the effect size, to find out how much heterogeneity remains after allowing for previously established reasons for heterogeneity. In practice, it is unlikely that all studies in a systematic review will have the necessary information to do a complete meta-regression including all previously documented reasons for heterogeneity. This could then be done on subsets of studies having the relevant information.

Heterogeneity is often measured with the I^2 statistic which varies between 0 and 100%, where 0% would indicate no heterogeneity and 100% large heterogeneity. Because the I^2 statistic is a relative measure, it is difficult to make a judgement about the absolute amount of heterogeneity. As a result, the use of the prediction interval has been suggested *(8-10)*.

The prediction interval provides an estimate of the distribution of the true effect sizes. To prevent overstating heterogeneity in observational studies, an 80% interval, and not the usual 95% interval, was chosen. For an 80% prediction interval, the true effect size for 80% of all populations would fall in this interval. This tells if the effect is consistent or if it varies substantially. It also tells if the effect is harmful in all populations, or if there is no effect in some populations or maybe even a preventive effect.

To make a judgement about the amount of heterogeneity that cannot be explained and that would be a reason for concern and a reason for downgrading, the following approach is proposed.
If the 80% prediction interval for a specific meta-analysis of relative risks is of the same size as the confidence interval, this indicates that there is no more variation in effect sizes than the statistical uncertainty. Then there is no reason for concern about heterogeneity.

However, if the prediction interval is considerably wider than the confidence interval (e.g., double the size) and overlaps with 1, there is reason for concern about heterogeneity. The effect sizes of the studies vary so much that with different samples of studies the conclusions of the meta-analysis could be substantially different. For example, an alternative conclusion could be that there would be no risk. In this case, the certainty of the body of evidence would be downgraded with one level.

Assessors need to provide a rationale for downgrading or not downgrading by explicitly addressing all of the issues mentioned above. This includes an assessment of how much of the heterogeneity can be explained.

Imprecision: downgrade one or two levels

Precision of the pooled effect size is another domain to be judged for downgrading. If there are only a few participants and the confidence interval around the pooled effect size is wide, one is less inclined to believe that the results reflect the true effects. If there is considerable imprecision, there is a reason to downgrade.

The cut-offs for downgrading because of imprecision given by the standard GRADE approach are applicable to clinical decision-making. Since in environmental health there are no clinical decision thresholds involved, only the second criterion of optimal information size can be applied to air pollution and health studies.

Therefore, the proposed approach consists of calculating the number of participants needed for a single study that can measure the relative risk of interest with sufficient precision *(11)*. If the

number of participants in the meta-analysis is considerably lower than the number that would be needed for an adequately powered study, the certainty of the evidence is rated down. This is a relatively conservative approach, and implies that the information size of the meta-analysis would need to be larger than the single study because heterogeneity has to be taken into account.

A method of calculating the sample size needed for a study with a specific relative risk and confidence interval was recently proposed by Rothman and Greenland *(12)*. As guidance, the calculation of the sample size needed to be able to assess a relative risk for mortality of 1.05 per 10 $\mu g/m^3$ increase of $PM_{2.5}$ with a confidence interval with a width of 0.09 (1.01−1.10) is provided below.

The event rate of mortality would be 0.0116 per person-year as in Ostro et al., 2010 *(13)*. This would lead to a number of about 940,000 person-years in the meta-analysis, containing sufficient information to assess the relative risk of interest with sufficient precision.

The event rate in the example above was observed over a five-year follow-up period in a cohort of female public school teachers aged around 54 years on average at baseline. As the confidence interval of the relative risk depends also and strongly on the event rate, the calculated number of about 940,000 person years should be viewed as indicative. It could be considerably smaller in older populations with higher event rates, and considerably larger in populations with lower event rates.

Separate calculations are needed for short-term studies which do not deal with person years but with numbers of daily events.

Publication bias: downgrade one level

Publication bias is assessed by a funnel plot and Egger's test. If the funnel plot upon visual inspection shows that small studies with non-harmful effects are missing, this would be an indication of publication bias. This means that small (imprecise) studies that have a relative risk smaller than 1 are missing. If there is no indication for these missing studies in the funnel plot, there is no use for the Egger's test, because significance will result from other factors causing heterogeneity *(10)*. The Egger's test would just be used to confirm suspected publication bias detected from the funnel plot.

It is important to note that the Egger's test can easily produce statistical significance for other reasons than publication bias in case of heterogeneity. Members of the Working Group noted that the Egger's test should not be used in case of heterogeneity, and that funnel plots should only include the studies included in the meta-analysis. Then, assessors should examine if small imprecise studies are missing in the funnel plots.

Other approaches to assessing reporting bias, such as a subgroup analysis of multi-centre studies compared to single city studies in case of evidence based on time series studies, an analysis of differences in effect estimates from earlier versus later studies, and a comparison to published results of attempts to quantify the magnitude of reporting bias, may help make a judgement.

Reasons for upgrading

The majority of the Working Group decided to recommend that upgrades for reasons of large effect size, all plausible confounding moving the relative risk estimate towards the null, and concentration–response gradient should be addressed independently from the results of applying the downgrading factors. Domains would be treated equally and independently, thus, leading to upgrading, downgrading or not changing the evidence level. A downgrade for any reason would not necessarily preclude upgrading for another reason.

Large magnitude of effect size: upgrade one level

The standard GRADE approach proposes upgrading the certainty of the evidence in observational studies if the pooled effect size is large or very large, so that 'the study design that is more prone to bias is unlikely to explain all of the apparent benefit or harm'. The cut-off point for a large effect size for harm is a relative risk > 2, while for a very large effect size is a relative risk > 5 *(13)*. These numbers are somehow arbitrary, and are not in the order of magnitude of the many relative risks reported in environmental health.

Instead of taking a certain value of the relative risk as the cut-off point, it is reasonable to judge whether confounding could have easily influenced the pooled effect size found in the meta-analysis. To this end, the application of the E-value approach is helpful *(14-17)*. This statistic is based on an assessment of how easily unmeasured confounders could explain away the relationship found between the exposure and the health outcome. It is based on the mathematical calculation of how large the effect of a confounder should be to explain away the relative risk that has been found in a study. With 'explain away', it is meant that such a confounder would reduce the relative risk that resulted from the observations in the study to 1. This effect (or E-value) is a function of the relative risk that has been found in a study or in a meta-analysis and is calculated as follows: E-value = RR + sqrt {RR * (RR – 1)}. The idea behind it is very similar to the 'large effect' concept in the standard GRADE framework but does not use absolute cut-offs for large effect sizes.

The judgement is then to ascertain if an unmeasured confounder could easily have an association with the exposure and the outcome with relative risks as large as or larger than the E-value. It is important to note that this is always the *covariate-adjusted* association between the unmeasured confounder and the outcome, and also the *covariate-adjusted* association between the unmeasured confounder and exposure to air pollution. If such a confounder could realistically have such strong relationships with both exposure and outcome, then unmeasured confounding could explain away the observed pooled relative risk. If one judges that it would be very unlikely that an unmeasured confounder would attain a relative risk as high as the E-value, then one can conclude that unmeasured confounding is unlikely to explain away the relative risk that has been observed. In that case, the certainty of the evidence can be upgraded because of a large effect size.

It is important to note that a major part of the judgement is what a realistic value for the relative risk of the unmeasured confounder could possibly be. Preferably, this should be based on what is known about strong confounders for the association at hand. For the association air pollution–mortality, smoking would be an obvious choice about which much information is available concerning its relationship with all-cause and cause-specific mortality. However, the residual association between smoking and air pollution is highly variable across published studies, and

calculations of E-values should report the covariate-adjusted associations with both air pollution and the outcome. The same logic applies to short-term studies where the covariate-adjusted associations between the confounder and the exposure (and the confounder–outcome) is relevant.

All plausible confounding shifts the relative risk towards the null: upgrade one level

Another proposed reason for upgrading is if all plausible confounding would shift the relative risk towards the null and still there would be a significant relative risk. This requires considerable judgement of possible confounders.

In most air quality and health studies, there would be a long list of possible confounders that would shift the relative risk in both directions. However, if one can reasonably argue that all confounding would have reduced the relative risk towards 1, then this will be a reason to upgrade the certainty of the evidence with one level.

Concentration–response gradient: upgrade one level

The standard GRADE proposes upgrading the certainty of the evidence if there is a concentration–response relationship between exposure and adverse health outcomes.

This domain is readily applicable to air quality and health studies. If there is an increase in risk with increasing exposure, either linearly or non-linearly, the certainty of the evidence would be upgraded with one level.

References

1. Schunemann H, Brożek J, Guyatt G, Oxman AD, editors. GRADE handbook for grading quality of evidence and strength of recommendations. Updated October 2013. Hamilton, ON: The GRADE Working Group; 2013 (https://gdt.gradepro.org/app/handbook/handbook.html, accessed 21 February 2020).
2. World Health Organization. WHO Handbook for Guideline Development, 2nd edition. Geneva: World Health Organization; 2014 (https://www.who.int/publications/guidelines/handbook_2nd_ed.pdf?ua=1; accessed 21 february 2020).
3. Morgan RL, Thayer KA, Bero L, Bruce N, Falck-Ytter Y, Ghersi D, et al. GRADE: Assessing the quality of evidence in environmental and occupational health. Environ Int. 2016;92-93:611-6. doi: 10.1016/j.envint.2016.01.004.
4. Woodruff TJ, Sutton P; Navigation Guide Work Group. An evidence-based medicine methodology to bridge the gap between clinical and environmental health sciences. Health Aff (Millwood). 2011;30(5):931-7. doi: 10.1377/hlthaff.2010.1219.

5. Hultcrantz M, Rind D, Akl EA, Treweek S, Mustafa RA, Iorio A, et al. The GRADE Working Group clarifies the construct of certainty of evidence. J Clin Epidemiol. 2017;87:4-13. doi: 10.1016/j.jclinepi.2017.05.006.

6. WHO Global Air Quality Guidelines Working Group on Risk of Bias Assessment. Risk of bias assessment instrument for systematic reviews informing WHO global air quality guidelines. Copenhagen: WHO Regional Office for Europe; 2020 (http://www.euro.who.int/__data/assets/pdf_file/0007/425518/Risk-of-bias-assessment-instrument-for-systematic-reviews-informing-who-global-air-quality-guidelines.pdf?ua=1, accessed 21 February 2020).

7. Guyatt GH, Oxman AD, Kunz R, Woodcock J, Brozek J, Helfand M, et al. GRADE guidelines: 8. Rating the quality of evidence — indirectness. J Clin Epidemiol. 2011;64(12):1303-10. doi: 10.1016/j.jclinepi.2011.04.014.

8. IntHout J, Ioannidis JP, Rovers MM, Goeman JJ. Plea for routinely presenting prediction intervals in meta-analysis. BMJ Open. 2016;6(7):e010247. doi: 10.1136/bmjopen-2015-010247.

9. Borenstein M, Higgins JP, Hedges LV, Rothstein HR. Basics of meta-analysis: I2 is not an absolute measure of heterogeneity. Res Synth Methods. 2017;8(1):5-18. doi: 10.1002/jrsm.1230.

10. Borenstein M. Common mistakes in meta-analysis and how to avoid them. Englewood, NJ: Biostat, Inc; 2019.

11. Rothman KJ, Greenland S. Planning study size based on precision rather than power. Epidemiology. 2018;29(5):599-603. doi: 10.1097/EDE.0000000000000876.

12. Ostro B, Lipsett M, Reynolds P, Goldberg D, Hertz A, Garcia C, et al. Long-term exposure to constituents of fine particulate air pollution and mortality: results from the California Teachers Study. Environ Health Perspect. 2010;118(3):363-9. doi: 10.1289/ehp.0901181.

13. Guyatt GH, Oxman AD, Sultan S, Glasziou P, Akl EA, Alonso-Coello P, et al. GRADE guidelines: 9. Rating up the quality of evidence. J Clin Epidemiol. 2011;64(12):1311-6. doi: 10.1016/j.jclinepi.2011.06.004.

14. VanderWeele TJ, Ding P. Sensitivity analysis in observational research: introducing the E-value. Ann Intern Med. 2017;167(4):268-274. doi: 10.7326/M16-2607.

15. Haneuse S, VanderWeele TJ, Arterburn D. Using the E-value to assess the potential effect of unmeasured confounding in observational studies. JAMA. 2019;321(6):602-603. doi: 10.1001/jama.2018.21554.

16. Ioannidis JPA, Tan YJ, Blum MR. Limitations and misinterpretations of E-values for sensitivity analyses of observational studies. Ann Intern Med. 2019;170(2):108-111. doi: 10.7326/M18-2159.

17. VanderWeele TJ, Mathur MB, Ding P. Correcting misinterpretations of the E-value. Ann Intern Med. 2019;170(2):131-132. doi: 10.7326/M18-3112.