# Chapter 6. Incorporating bias assessments into evidence synthesis

CHAPTER 6.

# Incorporating bias assessments into evidence synthesis

*Amy Berrington de González, Nathan DeBono, Alexander P. Keil, Deborah A. Lawlor, Ruth M. Lunn, and David A. Savitz*

## 6.1 Introduction

In the *IARC Monographs* programme, and in many other situations, experts are asked to examine, evaluate, and interpret a body of research that will then be used to make a judgement that could inform an authoritative statement, influence regulations, guide individual behaviours, or have other societal impact. In the *IARC Monographs* assessment process, the focus is on potentially preventable causes of cancer, but the same principles are applicable to other disease determinants and health outcomes. In public health, the determination of causation is rarely a simple yes–no decision. Rather, it requires the careful assembly of evidence and the use of inferential methods to reach a conclusion. Studies can provide information on the statistical relation between exposure and disease; by combining subject-matter expertise with an understanding of the study design and methods, considering complementary lines of research, and carefully examining the results, an assessment is made of the validity of the observed associations and their implications for inferences about causality. In almost all situations of interest, there will be more than one contributory study. The goal is to assess first the information value of each study, methodically and accurately, and then the totality of the available studies.

In this chapter, approaches are outlined for incorporating the wide array of bias assessment methods described in this book into the review process and evidence synthesis. This includes developing the process for the systematic review of key biases in individual studies and incorporating the bias assessment into the evidence synthesis. Two somewhat distinct approaches to the systematic review of biases are currently in use, which can be labelled as triangulation and algorithms. These two approaches are first described and contrasted, and then the rationale for a proposed third way is provided, drawing on the strengths of each. Three main steps in the bias-review process are outlined and illustrated with examples from the exposures used throughout this book. The chapter concludes with some discussion of methods for evaluating multiple sources of bias within a single study.

## 6.2 Frameworks for incorporating bias assessment into evidence synthesis

### 6.2.1 Triangulation

The triangulation of evidence from cancer epidemiology, animal bioassays, and mechanistic research is the overarching framework for the *IARC Monographs* review and classification system, as detailed in Chapter 1. Triangulation was introduced conceptually in Chapters 3 and 5 as a means of examining biases (specifically, confounding and selection bias) in individual studies. Triangulation can also serve as a framework for bias assessment across the epidemiological data. This approach emphasizes the benefits of examining the complete array of evidence to determine whether the varying strengths and limitations of the studies provide complementary information that helps in making an integrated assessment (Lawlor et al., 2016). The concept is particularly applicable when there is an array of studies with varying methodological strengths and limitations that could lead to bias in opposing directions.

Specifically, the aim in triangulation is to identify study designs (or approaches) that would be expected to have biases in opposing directions, to infer what a third, hypothetical, group of idealized studies would find. This inferred ideal can provide additional information about the probable bounds of a true causal effect. In practice, this can be implemented by contrasting studies through stratified meta-analysis or stratified forest plots. The approach requires consideration of the direction of the potential biases; this is an important strength. Example subgroups of studies that could be contrasted include the exposure setting, which might relate to the exposure level and degree of measurement error, for example studies of occupational versus environmental levels of exposure, cohort versus case–control study designs for assessment of recall bias, or cancer incidence versus mortality end-points for outcome misclassification. The study features should ideally involve complementary and exclusive biases that might affect one group of studies but not another. While no single study is likely to have all the desired positive features, a series of imperfect studies with complementary features could allow inference of what might be found in an ideal study.

By considering the full array of informative studies, there is an emphasis on corroboration, which links back to Hill's viewpoint on consistency of findings across a variety of locations and populations (Hill, 1965) and is consistent with the *IARC Monographs* Preamble (IARC, 2019), as noted in Chapter 1. Triangulation emphasizes the exploration of sources of heterogeneity. The reasoning is logical, intuitive, and flexible in being adaptable to a range of topics and diverse methodological concerns. These are all additional strengths. A weakness, currently, is that triangulation is a broad rather than a specific approach, lacking standardization; this could be invoked as a rationale that leads to a range of conclusions. In drawing on subject-matter expertise to interpret a given set of studies, there is latitude in what is emphasized and what is downplayed.

### 6.2.2 Algorithms

Several algorithms for bias assessment in epidemiology have been proposed, including Risk of Bias in Non-randomized Studies of Interventions (ROBINS-I) (Sterne et al., 2016) or of Exposure (ROBINS-E) (Higgins et al., 2022, 2024) and risk-of-bias scales, such as the Newcastle–Ottawa instrument (Deeks et al., 2003). A strength is that they offer a comprehensive set of rules and procedures to follow, with the intent of providing an evaluation that is replicable and objective and can be conducted by non-experts. A concern is the unwarranted degree of confidence that the algorithm gives the so-called correct answer (Igelström et al., 2021). There is no gold standard to know when an answer is right or wrong, and it is preferable to acknowledge the complexity of inferences about causality and accept the burden of explaining the reasoning that leads to the judgement, instead of simply invoking an algorithmic methodology. The aspiration of eliminating the subjectivity of reviewers and ensuring replicability is laudable, but it is unrealistic to expect that a generic algorithm for judging study quality will apply with equal validity to all exposures and all outcomes.

The comprehensive nature of the current algorithms, often involving a lengthy series of questions covering every potential source of bias, can also be a weakness. If there is no initial evaluation by subject-matter experts of the domains that are key or influential biases for the exposure and outcome of interest, then the application of the algorithm to every study tends to pare down the evidence that is used, with studies accepted or rejected due to possibly

minor or misplaced concerns rather than acknowledging that each has strengths and limitations. Many algorithms also do not emphasize an evaluation of the direction of the bias. In hazard identification, the direction is especially important. There is the potential for substantial loss of information about a potential hazard if all positive studies with bias towards the null were excluded, for example. The ability to assess consistency and the role of chance is also reduced if only a small subset of studies is retained.

These algorithms are often used in conjunction with the Grading of Recommendations, Assessment, Development, and Evaluations (GRADE) framework (Guyatt et al., 2008), developed for assessing clinical or other forms of experimental research, which automatically downgrades the value of observational studies in the evidence synthesis. Randomization is rarely ethical or feasible with etiological studies of cancer (other than prevention trials) and often requires the forfeit of other important study attributes, including exposure range, prolonged exposures, and study size. The strengths and weaknesses of different study designs will depend on the specific exposure and outcome under consideration.

Finally, although an algorithm may be presented as well-defined and systematic, there is still abundant opportunity to have the opinions of those implementing it influence the outcome. To the extent that there is a need for subject-matter expertise and an inherent intrusion of individual judgements, it is preferable to present the fact transparently rather than to mask it behind an algorithm.

### 6.2.3 Concluding thoughts about frameworks

There are strengths and weaknesses of triangulation and algorithms, as currently proposed, as bias assessment frameworks for epidemiological studies. A third way lies between the rigid approach of algorithms and the general approach of triangulation. This third way involves laying out a bias assessment process for the specific exposure and outcome under review that uses the full array of informative studies and the wide array of tools described in this book to assess the direction and magnitude of potential biases. This proposed approach is consistent with the review methods described in the *IARC Monographs* Preamble (IARC, 2019; see Chapter 1), which calls for Working Groups to integrate studies in evidence synthesis on the basis of their quality and informativeness but recommends against the use of checklists to assess biases and sources of error. It is also recommended that the bias assessment process be led by subject-matter experts, including epidemiologists, statisticians, and exposure assessors, again consistent with the *IARC Monographs* assessments. The following sections outline the key steps in this proposed approach and illustrate it with examples.

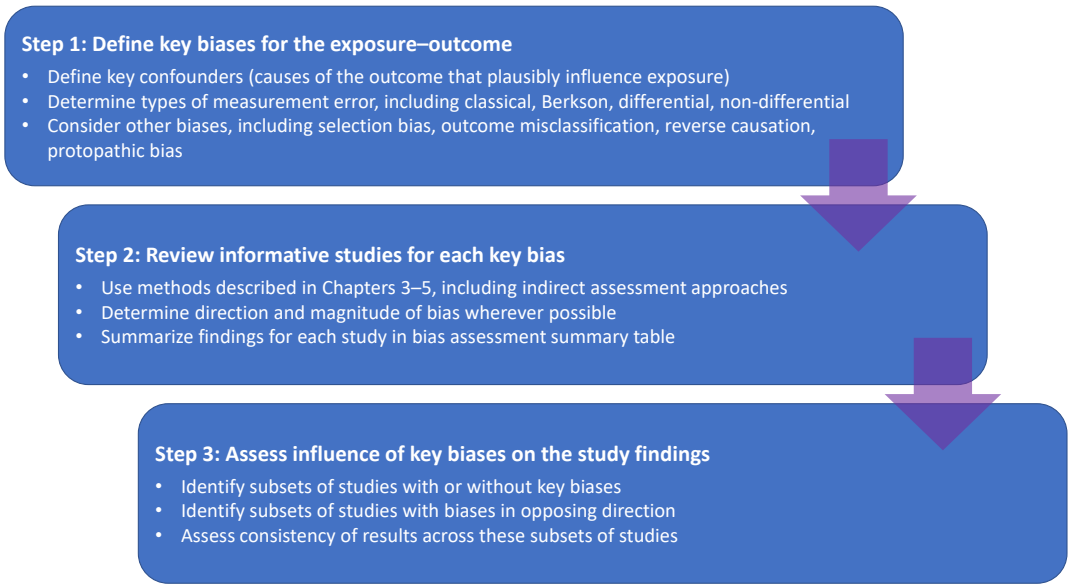
## 6.3 Developing the bias-review process

A bias-review process, developed by subject-matter experts, can guide the systematic review of biases in each individual study and at the evidence synthesis stage. There are typically several steps in the process, as outlined in Fig. 6.1: (i) a definition of the key biases for the exposure–outcome under consideration; (ii) a review, and a summary, of the informative studies for these key biases; and (iii) an assessment of the influence of the key biases on the study findings. The process is specific to each pair of exposures and outcomes under consideration and can be iterative. For the *IARC Monographs* evaluations, the Preamble and instructions for authors provide a starting point, and substance-specific issues can be added to the meeting-specific instructions for authors (IARC, 2024). These steps are described in more detail next.

### 6.3.1 Determining the key types of bias

A key step for the expert review group is to consider which of the many potential biases are of greatest concern. This will depend on the specific exposure–cancer outcome pair under review, and on the types of study that are available. A directed acyclic graph (DAG), as described in Chapter 2, can help the expert reviewers to reach agreement on the possible bias domains. Once these bias domains are agreed on, some specific signalling questions can be developed to guide the reviewers in their considerations. These questions should help identify the direction and likely magnitude of the bias, not simply its presence or absence. Chapters 3, 4, and 5 can help the reviewers make these determinations. Deciding which biases are not relevant, or not likely to be material, helps to focus the reviewers' attention on the critical subset. This process is illustrated in Examples 6.1 and 6.2.

**Fig. 6.1.** Steps in the bias-review process.




### **Example 6.1.** Selection of key biases for night shift work

Because night shift work is a complex exposure scenario, the *IARC Monographs* Working Group stated in its assessment of the evidence in humans that "exposure assessment quality of night shift work was a key parameter for the evaluation of the studies" (IARC, 2020), and the reviewers conducted an extensive evaluation of this aspect of each study. In contrast, the Working Group noted that although differences in lifestyle factors exist between day and night shift workers, these differences are usually small; this suggests that the reviewers considered confounding to be of lesser concern. Because there were many informative case–control studies, which tended to have more detailed exposure assessment, selection bias was examined, along with recall bias. (text continues on page 162)

### **Example 6.2.** Selection of key biases for opium consumption

There were a wide range of concerns about potential biases in the epidemiological studies of opium consumption, and the *IARC Monographs* Working Group documented its considerations in an annex to *IARC Monographs* Volume 126 (IARC, 2021), which serves as an example of a bias assessment framework. The Working Group noted that key potential biases for the examined studies of opium consumption included reverse causation (consumption of opium because of a cancer diagnosis) and protopathic bias (consumption of opium to alleviate prediagnostic symptoms). In addition, there were concerns about selection bias because there were several hospital-based case–control studies. Non-differential exposure misclassification and inclusion of infrequent opium users in the baseline category used for exposure–response analyses were thought to lead to downward bias. Finally, there were other strong risk factors for the cancers under study, particularly tobacco use, which had been shown in the exposure assessment review to be strongly related to opium use; thus, confounding was also a potential bias. (text continues on page 164)

### *(a) Guidance for identifying key confounders*

Once the key bias domains have been identified for the specific exposure–cancer scenario under investigation, the review team should provide additional details to guide the bias assessment. For confounding, the Working Group members should use their expertise and literature reviews to identify all the key confounders, i.e. those variables most likely to bias the effect estimate and distort its interpretation if they are not controlled for in the study. The use of DAGs can guide and help document these decisions (see Chapter 2). An approach to this identification is given in Side Box 6.1 and Example 6.3. The methods in Chapter 3 can help in assessing the likely direction and magnitude of confounding. There should also be consideration of whether certain variables could be effect modifiers or mediators, rather than confounders, because adjustment for these could introduce, rather than remove, bias (see Chapter 2 for more details).

### *(b) Guidance for assessing misclassification and mismeasurement of exposure*

In general, the bias framework for exposure misclassification should cover how well the exposure proxy approximates the exposure of interest, the extent of measurement error, and whether the measurement error is differential or non-differential. Side Box 6.2 lists scoping questions to inform the bias evaluation. The methods described in Chapter 4 can help to determine the likely direction and possible magnitude of bias from misclassification and mismeasurement of exposure, as illustrated in Example 6.4.

### *(c) Guidance for assessing other key biases*

The detailed guidance described above for confounding and measurement error provides examples of thorough assessment of the key concerns for these topics. Other topics may call for analogous assessments of other types of bias, for example selection bias, healthy worker effects, and outcome misclassification. For each key bias, a set of questions should be identified and guidance provided. For example, for selection bias, reviewers should consider sources of bias such as study inclusion or exclusion criteria, sources of control participants, and missing data or loss to follow-up (Example 6.5). It may help to use DAGs to illustrate sources of selection bias, including colliders. For a detailed evaluation of how to identify selection bias in case–control studies, see Section 5.3.

## *6.3.2 Summarizing the bias assessment and synthesizing across studies*

A table summarizing the results from the review of the key biases in each study is recommended. For instance, in Example 6.6, for an analysis of studies on opium consumption and

**Side Box 6.1.** Approach for identifying key confounders

(i) Identify the known causes of the cancer (e.g. those with *sufficient* or *limited* evidence of causality) by consulting experts with relevant subject knowledge and using authoritative sources, such as the *IARC Monographs* and the *IARC Handbooks of Cancer Prevention*, the United States National Toxicology Program Report on Carcinogens, and the World Cancer Research Fund. Specify the hypothesized direction of the confounder–cancer association (e.g. relative risk [RR] > 1 or RR < 1).

(ii) Identify which cancer causes are plausibly related to the exposure of interest, by using authoritative sources and consulting experts with relevant subject knowledge. This information is often reported in the section on exposure characterization of the relevant *IARC Monograph*. Specify the hypothesized direction of the confounder–exposure association (e.g. RR > 1 or RR < 1).

(iii) Research (e.g. conduct literature searches, seek expert opinion on mechanistic data) whether the identified potential confounders could be mediatory (in the causal pathway between the exposure and cancer) rather than confounders. It may be helpful to construct a DAG to identify mediators and colliders, which should not be controlled for in studies.

(iv) Identify the minimal set of key variables necessary to control for confounding, and assess the expected direction of the bias (the methods outlined in Chapters 2 and 3 can be helpful). (text continues above)

### **Example 6.3.** Specifying key confounders

Returning to the example of night shift work in relation to breast cancer, this example illustrates how the approach outlined in Side Box 6.1 can be used to specify key confounders.

Table 6.1 lists the causes of female breast cancer identified from IARC, the World Cancer Research Fund, the United States National Toxicology Program Report on Carcinogens, and literature reviews, and the subset of these that could be considered as potential key confounders for night shift work. Age at first full-term pregnancy could be considered the key confounder for reproductive breast cancer factors because other factors, such as parity, are often related to it, and some of their confounding effects are likely to be controlled for by controlling for age at first full-term pregnancy. Other pharmacological and lifestyle factors, such as the use of oral contraceptives and tobacco smoking, might not be key confounders because of relatively weak associations with breast cancer. In contrast, although a family history of breast cancer is strongly associated with breast cancer risk, it would be unlikely to be associated with night shift work and would therefore not be a key confounder. Occupational exposure to ionizing (cosmic) radiation could be a key confounder in flight crew studies because of its high correlation with night work hours. (text continues on page 164)

**Table 6.1.** Potential key confounders for night shift work and female breast cancer

| Potential confounding factors | Causes of female breast cancer[a] | Key confounders (and expected directions) |
|---|---|---|
| Reproductive and family history factors | Early age at menarche, late age at first full-term pregnancy, nulliparity, menopausal status or age at menopause, no breastfeeding, family history of breast cancer | Young age at first full-term pregnancy or parity. These are protective for breast cancer and are probably negatively associated with night shift work; therefore, confounding away from the null. |
| Lifestyle factors | Lack of physical activity (primarily postmenopausal breast cancer), obesity (increases risk in postmenopausal women; decreases risk in premenopausal women), consumption of alcoholic beverages, tobacco smoking | Obesity is a risk factor for breast cancer and is probably positively associated with night shift work; therefore, bias probably away from the null. Note that obesity could be a mediator or a confounder (or both). |
| Pharmacological factors | Diethylstilbestrol, estrogen–progestogen contraceptives, hormone menopausal therapy (estrogen–progestogen or estrogen only), digoxin | |
| Demographics | Age, socioeconomic status, education level | Age, socioeconomic status, and education level are negatively associated with breast cancer, but the direction of the association with night shift work depends on the profession. |
| Occupational agents | X-radiation, gamma radiation, ethylene oxide, polychlorinated biphenyls | Cosmic radiation for aircrew workers. The direction of the bias depends on the comparison group (e.g. day workers or non-workers) and the study population. |

[a] NTP (2018).

**Side Box 6.2.** Approach for assessing exposure misclassification and measurement error

- How was exposure assessed in the studies under review (e.g. questions, records, environmental measurements, biomarkers)?
- The Working Group should research the following questions for each type of exposure assessment.
  - What was the temporal sequence of exposure and outcome measurement? Could disease status have affected the exposure measurement?
  - Are there methods that are prone to major error or are biologically inappropriate for the exposure of interest? For example, a biomarker with a short half-life might not be informative for evaluating cancer risk.
  - What are the ideal methods for evaluating exposure?
  - What are the potential sources of measurement error?
  - What is the type of measurement error (e.g. classical, Berkson, differential, non-differential)?
  - Are there validation studies available? What values of sensitivity and specificity do the validation studies report? (text continues on page 164)

**Example 6.4.** Assessing exposure misclassification

For studies on night shift work and breast cancer, the most common methods to assess and classify exposure involved using questionnaires, payroll records, or a population-based job-exposure matrix (e.g. based on survey data reporting the percentage of night shift workers for different job categories). Table 6.2 lists questions and considerations for assessing the potential biases from exposure misclassification in the studies on night shift work. (text continues on page 164)

**Table 6.2.** Assessment of exposure misclassification for studies on night shift work

| Questions | Guidance, comments |
|---|---|
| What is the source of the exposure assessment?<br>Is there concern that the exposure assessment did not distinguish between exposed and non-exposed people or among exposure categories during a relevant time window of exposure?<br>What are the likely direction and magnitude of bias? | Questionnaires, interviews:<br>Ideally, the questionnaire should cover actual hours worked and lifetime work history. The group defined as unexposed might be exposed (i.e. previous night shift work) in studies ascertaining current exposure. Ideally, in cohort studies, information should be collected after baseline.<br>Job-exposure matrix (JEM):<br>Population-based JEMs are less informative than industry-specific JEMs. Information (e.g. census data) of the proportion of night shift workers in the same geographical region as the study population could provide some indication of the quality of the data.<br>Payroll records:<br>These are objective but are usually not complete because industry-specific records do not capture lifetime exposure from jobs in other workplaces. In general, bias from lower-quality exposure assessment is likely to be non-differential and towards the null (see Chapter 4 for exceptions). |
| Is there concern about differential recall? | Differential recall bias (most likely away from the null) is a potential concern in case–control studies. It may be less likely in studies published before the 2007 IARC evaluation of shift work (IARC, 2010).<br>External research may help inform the assessment of recall bias (see Chapter 5 for examples). |

**Example 6.5.** Identifying selection bias

Table 6.3 illustrates how this approach to assessing the potential for selection bias can be applied to a bias-review framework for case–control studies on opium consumption and various cancers. (text continues on page 164)

**Table 6.3.** Identifying selection bias for case–control studies on opium consumption

| Question | Guidance, comments |
|---|---|
| Is there concern that selection into (or out of) the study was related to both exposure and outcome, and what is the likely direction of the bias? | Hospital control participants:<br>Potential bias downwards if opium use is related to hospitalization and hospital control participants are more likely than the general population to have used opium. |
| | Neighbourhood control participants:<br>Potential bias upwards if control participants who use opium are less likely to participate (e.g. leading to a lower exposure prevalence). |

bladder cancer, Table 6.4 has one row per study and a column for each key bias, and gives the likely direction of potential bias. It is then easy to see the biases that have been identified for each study, and the groups of studies that have been identified with a certain bias. The table can be used to inform a triangulation process by identifying subsets of studies with differing key sources of biases, particularly where some studies would be expected to produce bias in opposing directions. The table also shows whether biases cluster within subsets of studies; this might make it difficult to separate the impact of specific biases. When multiple key biases affect a study, assessment of the total (resultant) bias is non-trivial. Section 6.4 describes some approaches and the related challenges involved in assessing multiple biases within a single study.

The extent to which it is then feasible to integrate study results and bias assessment can be influenced by how many informative human studies have been identified for review. For triangulation, the aim is to compare results from at least two, but ideally more, studies that have different key sources of bias. The study results can be contrasted via stratified forest plots or, more formally, by means of stratified meta-analysis to explore the impact of the bias. When there are many informative studies, the opportunities for bias assessment through triangulation are increased, particularly if there is a variety of settings and study designs. The different steps in the process are illustrated with examples including studies on opium consumption and bladder cancer (Example 6.6), a situation with only a few informative studies (mobile phone use and glioma; Example 6.7), and a quantitative triangulation of meta-analysis results where there are a large number of studies (red meat consumption and colorectal cancer; Example 6.8).

## 6.4 Methods for studying multiple biases

As seen in the examples throughout this book, studies could be subject to multiple key biases. At the evidence synthesis stage, the reviewers will then need to consider what the combined effect of those biases might be, and whether the combination could alter the interpretation. To answer this question requires consideration of the magnitude of each bias, along with the direction of the bias and some understanding of whether the biases act independently. This section discusses the issues that need to be considered when assessing the likely impact of multiple biases, how to approach multiple-bias sensitivity analyses, and when an individual-level data reanalysis could be important. Annex 3 includes a worked example of a formal multiple-bias analysis for a study on opium consumption and bladder cancer, which illustrates the complexity and the need to specify multiple parameters. Because of the

**Example 6.6.** Bias assessment summary table

In the review of the human evidence for *IARC Monographs* Volume 126 on opium consumption and bladder cancer, one cohort study and several case–control studies were considered informative (IARC, 2021). As noted in Section 6.3.1, there were a considerable number of potential key biases, which were discussed in an annex, titled "Methodological considerations for epidemiological studies on opium consumption and cancer" (IARC, 2021). A meta-analysis published subsequently used the bias assessment to explore between-study heterogeneity; that assessment is used here to illustrate how the biases can be summarized and synthesized (Miranda Filho et al., 2023).

Table 6.4 shows that most studies were not considered to be at risk of material (major) confounding bias or reverse causation, but that many of the case–control studies were considered to be at risk of selection bias and information bias. The direction of selection bias was identified as likely downwards in several hospital-based case–control studies, but of uncertain direction in others. The potential for recall bias and exposure misclassification was considered quite low, but these biases could operate in different directions, hence the arrow showing that this could result in bias towards or away from the null. In all the studies, a positive association was found between opium consumption and bladder cancer, but the magnitude of risk for ever or never having used opium varied widely, with an odds ratio of 2.47 to 8.23 and a summary estimate of 4.07 (95% confidence interval [CI], 3.23–5.12). Miranda Filho et al. (2023) conducted sensitivity analyses by excluding studies with various biases (e.g. selection bias, information bias). The summary relative risk was slightly lower in the studies considered to have low risk of selection bias (odds ratio [OR], 3.40; 95% CI, 2.70–4.30) or information bias (OR, 3.69; 95% CI, 3.01–4.41) but was still strongly supportive of a positive association.

**Table 6.4.** Bias assessment summary for studies on opium consumption (ever vs never use) and bladder cancer based on major concerns, as defined and identified by Miranda Filho et al. (2023)[a]

| Study (first author) | OR or RR (CI)[b] | Design | Confounding | Reverse causation | Selection bias | Information bias | Protopathic bias |
|---|---|---|---|---|---|---|---|
| Sheikh | 2.86 (1.47–5.56) | co | | | | | |
| Aliasgari | 2.60 (0.80–8.47) | c–c(h) | | | ← | ↔ | |
| Aliramaji | 4.10 (1.59–10.55) | c–c(h) | | | ← | ↔ | |
| Sadeghi | 2.70 (0.18–40.81) | c–c(h) | | | ← | ↔ | |
| Nourbakhsh | 3.87 (1.98–7.57) | c–c | | | ↔ | ↔ | |
| Tootoonchi | 2.45 (0.98–6.14) | c–c | | | ↔ | ↔ | |
| Abdolahinia | 8.23 (3.82–17.71) | c–c | | | ↔ | ↔ | |
| Akbari | 3.90 (1.28–11.85) | c–c | | | | | |
| Hadji | 3.40 (2.69–4.29) | c–c | | | | | |
| Rashidian | 4.40 (2.94–6.59) | c–c | | | | | |
| Ghadimi | 4.96 (1.07–22.96) | c–c(h) | | | ← | ↔ | |
| Hosseini | 4.16 (2.67–6.47) | c–c(h) | | | ← | | |
| Ketabchi | 7.99 (5.20–12.27) | c–c | | | ↔ | ↔ | |
| Lofti | 3.01 (1.73–5.23) | c–c | | | | ↔ | |
| Shakhssalim | 2.57 (1.55–4.26) | c–c | | | | ↔ | |

c–c, case–control; c–c(h), hospital-based case–control; CI, confidence interval; co, cohort; OR, odds ratio; RR, relative risk.

[a] Arrows indicate the direction of the biases: ←, downwards; ↔, uncertain direction. Blank indicates that the reviewers concluded that there was no substantial bias.

[b] Controlling for tobacco smoking, where available.

**Example 6.6.** Bias assessment summary table (continued)

The Working Group concluded that chance, bias, and confounding could be ruled out because of the strong associations and the consistency across studies and across the study designs (e.g. the cohort and the case–control studies with different sources of control participants). The Working Group did not use the term *triangulation* but commented, "It is notable that the results of all studies, regardless of design, point in the same direction" (IARC, 2021). (text continues on page 167)

**Example 6.7.** Bias assessment summary with few informative studies

In a review of studies on radiofrequency electromagnetic field radiation exposure (mainly through mobile phone use) and brain tumours (IARC, 2013), the reviewers considered most of the early small case–control studies to be relatively uninformative. Therefore, evaluation of the human evidence was based largely on two large case–control studies: the Interphone multicentre case–control study (Cardis et al., 2011) and a large case–control study in Sweden (Hardell et al., 2011).

In the case–control study in Sweden, with 1148 cases of glioma and 2438 control participants, Hardell et al. (2011) reported a monotonically increasing risk of glioma with increasing cumulative duration of mobile phone use, with an odds ratio of 3.2 (95% CI, 2.0–5.1) for > 2000 hours use compared with no use. In the Interphone study, with 2708 cases of glioma and 2792 control participants, cumulative call time was divided into deciles, with a referent comprising those who had never regularly used mobile phones. In contrast to the findings from the case–control study in Sweden, in the Interphone study, the odds ratios were mostly < 1 (ranging from 0.7 to 1.05), except for the highest category, of ≥ 1640 hours of cumulative call time (OR, 1.40; 95% CI, 1.03–1.89). Because these were case–control studies based on self-reported mobile phone use, the review group identified differential measurement error (recall bias) and selection bias as the key potential sources of bias. Because there are few established risk factors for glioma, confounding was considered less of an identifiable problem.

Selection bias was of greater potential concern in the Interphone study, because the participation rates were relatively low, especially for control participants (64% for cases and 53% for controls). In the case–control study in Sweden, participation rates were higher and non-differential (85% for case participants and 84% for control participants). In the Interphone study, a short non-response questionnaire revealed that the participation rate was higher in regular mobile phone users, particularly for case participants. When the analysis was restricted to regular users (i.e. by changing the reference category), the odds ratios for cumulative call time changed qualitatively to become mostly > 1 (increasing by 20–50%). Although there was still no clear evidence of a dose–response relation across the 10 categories of duration, the odds ratio for ≥ 1640 hours of cumulative call time increased from 1.40 (95% CI, 1.03–1.89) to 1.82 (95% CI, 1.15–2.89).

There were also extensive efforts to evaluate the quality of the exposure data in the Interphone study; these included a substudy with software-modified phones and phone records, which found substantial reporting error, with some indication of greater overreporting by case participants (Vrijheid et al., 2006, 2009). Exclusion of all participants who reported usage for > 5 hours per day decreased the odds ratio in the highest decile from 1.40 to 1.27 (95% CI, 0.92–1.74), but truncation at 5 hours per day did not influence the odds ratio. As explained in Chapter 4, bias from non-differential misclassification in categorical variables is not necessarily towards the null. Table 6.5 summarizes the bias assessment for the key domains, and the likely direction of the bias for the two informative studies.

**Example 6.7.** Bias assessment summary with few informative studies (continued)

**Table 6.5.** Bias assessment summary for case–control studies on mobile phone use and glioma[a]

| Study | Risk estimate (95% CI) | Information bias | Selection bias |
|---|---|---|---|
| Case–control study in Sweden (Hardell et al., 2011) | 3.2 (2.0–5.1)[b] | ↔ | |
| Interphone study (Cardis et al., 2011) | 1.40 (1.03–1.89)[c] | ↔ | ← |

CI, confidence interval.

[a] Arrows indicate the direction of the biases: ←, downwards; ↔, uncertain direction. Blank indicates that the reviewers concluded that there was no substantial bias.

[b] Highest exposure category of > 2000 hours of cumulative call time.

[c] Highest exposure category of ≥ 1640 hours of cumulative call time.

Because there were only two informative studies and they had a similar design (population-based case–control studies with self-reported mobile phone use), triangulation was not possible. The higher risk of selection bias in the Interphone study, with some evidence that this was biased downwards, could partly explain the difference in the magnitude of the risk estimates for the highest exposure category. However, these studies share the limitation of potential for recall bias and exposure misclassification, which could have opposing directions. Therefore, the assessment of the human evidence by the committee was that although there was a positive association between mobile phone use and the incidence of glioma, chance, bias, or confounding could not be ruled out with reasonable confidence. Multiple-bias analysis could have been used to further explore the combined effect of these biases. As noted in Section 6.4, this is a complex task and involves several assumptions and specification of multiple-bias parameters but can provide bounds on the plausible range of results. (text continues on page 167)

effort involved, it is worth considering whether multiple-bias assessment is necessary. For example, if the study result is positive and all key biases are expected to be towards the null, it is unnecessary to carry out a formal multiple-bias analysis for hazard identification.

There are two different approaches for sensitivity analysis: bias-level sensitivity analysis and target-adjusted sensitivity analysis. In bias-level sensitivity analysis, plausible bias values and structures are used to identify a range of results that the study could have obtained. When dealing with multiple biases, the order of corrections must be considered. For example, should one adjust for confounding or exposure misclassification first? Fox et al. (2021) recommend what they term sequential bias analysis, in which biases are adjusted for sequentially in the reverse order of which they likely occurred. A common sequence in which biases arise would be confounding, followed by selection bias, and finally exposure misclassification, but this is not always the case. The order of analysis matters because sensitivity and specificity parameters, for example, may differ, depending on whether misclassification of the exposure or outcome occurs before or after study selection (Example 6.9). Ross et al. (2022) show how adjusting for biases in the wrong order using individual-level data can lead to misadjustment and residual bias.

Example 6.9 highlights the challenges of conducting a multiple-bias analysis, of which there are very few examples in the literature. If evidence hinges on a single study in which multiple biases are suspected, such an analysis may be informative, but it should be interpreted cautiously, because of the inherent dependence on the accuracy of bias parameters. Probabilistic bias analysis accounts for uncertainty in the bias parameters and is discussed in Chapters 4 and 5. This uncertainty is quantified by proposing a distribution, rather than a single value, for each bias parameter. At the extreme end of probabilistic bias analysis is bounding, which involves finding the largest amount of bias that could result from the plausible distribution of bias parameters. In principle, a bounding approach can help to answer questions about

**Example 6.8.** Bias assessment summary using triangulation

This example is an illustration of triangulation using meta-analyses of studies of the association between red meat consumption and colorectal cancer. Results are stratified according to study designs that are likely to have biases in opposing directions: cohort and case–control. It is assumed that non-differential exposure misclassification is a source of bias towards the null in cohort studies with a single dietary questionnaire of limited detail. Also, it is assumed that recall bias is away from the null in the case–control studies, for example through case participants overreporting their exposure because of their diagnosis. These biases are unrelated, in that each bias affects one group of studies (i.e. cohort studies, case–control studies) but not the other.

From the results reported by Norat et al. (2002), the meta-effect estimate (Table 6.6) for the highest versus lowest quantile of consumption from the cohort studies (1.27) is slightly lower than that from the case–control studies (1.36). It is then possible to make inferences about a third, hypothetical, meta-effect estimate from an idealized study with no biases. Triangulation of the stratum-specific effect estimates suggests that the true causal effect may be between these two values. In this way, a bounded range of the magnitude of the causal effect is obtained, using information from two groups of studies. This approach is likely to be more informative than making inferences from one group of studies, because of a perceived methodological strength, while ignoring another. The IARC Working Group also identified several key confounders for the association between red meat consumption and colorectal cancer, including total energy (caloric) intake, physical activity, smoking, and body mass index. Stratified meta-analyses based on the degree of control for confounding within the subsets of case–control and cohort studies could provide further insight into the impact of confounding, and potential mediation for body mass index (as discussed in Example 2.1a). Additional insights into the potential impact of measurement error in red meat consumption are also shown in Example 4.22, which illustrates regression calibration. Calibration corrections of this type can be important in meta-analyses because they can reduce an important source of heterogeneity in effect estimates. (text continues on page 167)

**Table 6.6.** Example triangulation exercise, comparing meta-analysis results from studies of the association between red meat consumption and colorectal cancer

| Strata | Source of bias | Direction of bias | Number of studies[a] | Meta-effect estimate (95% CI)[a] | Triangulated meta-effect estimate |
|---|---|---|---|---|---|
| Cohort studies | Non-differential exposure misclassification | Towards the null | 9 | 1.27 (1.11–1.45) | 1.27–1.36 |
| Case–control studies | Recall bias | Away from the null | 14 | 1.36 (1.17–1.59) | |

CI, confidence interval.

[a] Results from Norat et al. (2002) for the highest quantile of red meat consumption.

**Example 6.9.** Multiple-bias analysis

A multiple-bias analysis within a systematic review may be most usefully undertaken when one has to consider a study of moderate to large size and evidence is uncertain. The study by Aliramaji et al. (2015) was one of the largest conducted to examine the relation between opium consumption and bladder cancer, and it was suspected to suffer from multiple key biases, whose directions might have offset each other (Table 6.4). In the original study publication, a crude odds ratio of 2.7 was reported for the opium consumption–bladder cancer association. Multiple biases were likely in this study. First, frequency matching on sex without adjustment would have introduced selection bias. Second, there was concern about exposure misclassification, because of the illicit nature of opium use. Third, there was a potential for uncontrolled confounding by smoking and sex (which are not noted in Table 6.4 because adjusted estimates were used for that determination, rather than the crude estimate reported by Aliramaji et al., 2015). Bias parameters to adjust for each of these biases were drawn from various sources, including survey data and a validation study of recent opium use conducted for a hospital-based cohort. A limitation of the bias analyses is the lack of validation studies of long-term opium use, which would have yielded misclassification parameters for long-term use. Instead, bias parameters were drawn from studies of recent use, which were available because there are reliable biomarkers of recent opium exposure.

The use of matching on sex probably led to downward selection bias because it resulted in an oversampling of men, who were less likely to be unexposed control participants. However, there was also uncontrolled confounding by sex, which was considered to be upwards. Adjustment for all three biases in the reverse order of which they were expected to occur (here the adjustment order was selection bias, exposure misclassification, and confounding) yielded an adjusted summary odds ratio of 8.6, suggesting that the bias in the study by Aliramaji et al. (2015), given the best available estimates of bias parameters, was downwards. This large change occurred because most adjustments were in the same direction. A full probabilistic bias analysis was not straightforward, given the studies from which the bias parameters were drawn. To address this partially, a sensitivity analysis was performed, in which the sensitivity parameters for exposure misclassification were varied within a relatively narrow range of plausible values. Even with this narrow range, the adjusted odds ratio ranged from 1.3 (no misclassification) to 11.2 (differential misclassification); this emphasizes the potential influence of this source of bias. Full details of this example are provided in Annex 3, along with R code (online only; available from: https://publications.iarc.who.int/634#supmat). (text continues on page 170)

whether an observed association might be due to bias alone. However, in practice, this bounding approach has not been widely used for multiple-bias analysis and is limited when considering biases that might offset each other. Finally, target-adjusted sensitivity analysis, such as the E-value (described in Section 3.3.4(c), which outlines a modified approach), involves identifying the extent of bias necessary for a given study result to be compatible with the null (or another) hypothesis. This approach would also likely be very difficult in a multiple-bias analysis using only published data and would involve unrealistic assumptions. Smith et al. (2021) give an example based on individual subject data.

## 6.5 Summary

This chapter provides pragmatic guidance on the development and application of bias assessment as part of the evidence synthesis process. A third way is offered, which lies between the rigid approach of algorithms and the general approach of triangulation. A critical philosophical distinction from the algorithmic approach is that this bias assessment should be developed and applied by multidisciplinary experts. This expertise facilitates the identification of key sources of bias for the specific exposure–cancer relation under review. Focusing on key sources of bias facilitates the review process and avoids the elimination of informative studies due to minor biases or biases that do not change the causal interpretation.

It is recommended to retain all informative studies and to document

the potential key biases, including their direction and, if possible, their magnitude. The array of studies can then be used to evaluate biases indirectly, and to triangulate epidemiological evidence by comparing results from subsets of studies with different key biases. The wide array of tools described in this book provides methods to evaluate the direction and magnitude of bias, drawing on external data where necessary. It is hoped that, in the future, these bias analyses will be incorporated into the results section in more original study publications, as outlined in Chapter 7, reducing the need for speculation in the ubiquitous paragraph on strengths and limitations in the discussion section of publications. This will strengthen the field of epidemiology and facilitate the bias assessment work of review teams.

# References


Aliramaji A, Kaseean A, Yousefnia Pasha YR, Shafi H, Kamali S, Safari M, et al. (2015). Age distribution types of bladder cancers and their relationship with opium consumption and smoking. Caspian J Intern Med. 6(2):82–6. PMID:26221505

Cardis E, Armstrong BK, Bowman JD, Giles GG, Hours M, Krewski D, et al. (2011). Risk of brain tumours in relation to estimated RF dose from mobile phones: results from five Interphone countries. Occup Environ Med. 68(9):631–40. doi:10.1136/oemed-2011-100155 PMID:21659469

Deeks JJ, Dinnes J, D'Amico R, Sowden AJ, Sakarovitch C, Song F, et al.; International Stroke Trial Collaborative Group; European Carotid Surgery Trial Collaborative Group (2003). Evaluating non-randomised intervention studies. Health Technol Assess. 7(27):iii–x, 1–173. doi:10.3310/hta7270 PMID:14499048

Fox MP, MacLehose R, Lash TL (2021). Applying quantitative bias analysis to epidemiologic data. Statistics for biology and health. 2nd ed. Cham, Switzerland: Springer. doi:10.1007/978-3-030-82673-4

Guyatt GH, Oxman AD, Vist GE, Kunz R, Falck-Ytter Y, Alonso-Coello P, et al.; GRADE Working Group (2008). GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. BMJ. 336(7650):924–6. doi:10.1136/bmj.39489.470347.AD PMID:18436948

Hardell L, Carlberg M, Hansson Mild K, Eriksson M (2011). Case-control study on the use of mobile and cordless phones and the risk for malignant melanoma in the head and neck region. Pathophysiology. 18(4):325–33. doi:10.1016/j.pathophys.2011.06.001 PMID:21764571

Higgins J, Morgan R, Rooney A, Taylor K, Thayer K, Silva R, et al. (ROBINS-E Development Group) (2022). Risk of bias in non-randomized studies - of exposure. Launch version, 1 June 2022. Available from: https://www.riskofbias.info/welcome/robins-e-tool.

Higgins JPT, Morgan RL, Rooney AA, Taylor KW, Thayer KA, Silva RA, et al. (2024). A tool to assess risk of bias in non-randomized follow-up studies of exposure effects (ROBINS-E). Environ Int. 186:108602. doi:10.1016/j.envint.2024.108602 PMID:38555664

Hill AB (1965). The environment and disease: association or causation? Proc R Soc Med. 58(5):295–300. doi:10.1177/003591576505800503 PMID:14283879

IARC (2010). Painting, firefighting, and shiftwork. IARC Monogr Eval Carcinog Risks Hum. 98:1–804. Available from: https://publications.iarc.who.int/116 PMID:21381544

IARC (2013). Non-ionizing radiation, Part 2: Radiofrequency electromagnetic fields. IARC Monogr Eval Carcinog Risks Hum. 102:1–462. Available from: https://publications.iarc.who.int/126 PMID:24772662

IARC (2019). Preamble to the *IARC Monographs* (amended January 2019). Lyon, France: International Agency for Research on Cancer. Available from: https://monographs.iarc.who.int/iarc-monographs-preamble-preamble-to-the-iarc-monographs/.

IARC (2020). Night shift work. IARC Monogr Identif Carcinog Hazard Hum. 124:1–371. Available from: https://publications.iarc.who.int/593 PMID:33656825

IARC (2021). Opium consumption. IARC Monogr Identif Carcinog Hazard Hum. 126:1–253. Available from: https://publications.iarc.who.int/600 PMID:36395294

IARC (2024). *IARC Monographs* Preamble – Instructions for authors. Lyon, France: International Agency for Research on Cancer. Available from: https://monographs.iarc.who.int/preamble-instructions-for-authors/.

Igelström E, Campbell M, Craig P, Katikireddi SV (2021). Cochrane's risk of bias tool for non-randomized studies (ROBINS-I) is frequently misapplied: a methodological systematic review. J Clin Epidemiol. 140:22–32. doi:10.1016/j.jclinepi.2021.08.022 PMID:34437948

Lawlor DA, Tilling K, Davey Smith G (2016). Triangulation in aetiological epidemiology. Int J Epidemiol. 45(6):1866–86.doi:10.1093/ije/dyw314 PMID:28108528

Miranda Filho A, Turner MC, Warnakulasuriya S, Richardson DB, Hosseini B, Kamangar F, et al. (2023). The carcinogenicity of opium consumption: a systematic review and meta-analysis. Eur J Epidemiol. doi:10.1007/s10654-023-00969-7 PMID:36773182

Norat T, Lukanova A, Ferrari P, Riboli E (2002). Meat consumption and colorectal cancer risk: dose-response meta-analysis of epidemiological studies. Int J Cancer. 98(2):241–56. doi:10.1002/ijc.10126 PMID:11857415

NTP (2018). Report on Carcinogens Protocol: night shift work and light at night (LAN): human cancer studies. Office of the Report on Carcinogens, Division of the National Toxicology Program, National Institute of Environmental Health Sciences, United States Department of Health and Human Services. Available from: https://ntp.niehs.nih.gov/ntp/roc/protocols/electric_light_508.pdf.

Ross RK, Breskin A, Breger TL, Westreich D (2022). Reflection on modern methods: combining weights for confounding and missing data. Int J Epidemiol. 51(2):679–84. doi:10.1093/ije/dyab205 PMID:34536004

Smith LH, Mathur MB, VanderWeele TJ (2021). Multiple-bias sensitivity analysis using bounds. Epidemiology. 32(5):625–34. doi:10.1097/EDE.0000000000001380 PMID:34224471

Sterne JAC, Hernán MA, Reeves BC, Savović J, Berkman ND, Viswanathan M, et al. (2016). ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. BMJ. 355:i4919. doi:10.1136/bmj.i4919 PMID:27733354

Vrijheid M, Cardis E, Armstrong BK, Auvinen A, Berg G, Blaasaas KG, et al.; Interphone Study Group (2006). Validation of short term recall of mobile phone use for the Interphone study. Occup Environ Med. 63(4):237–43. doi:10.1136/oem.2004.019281 PMID:16556742

Vrijheid M, Richardson L, Armstrong BK, Auvinen A, Berg G, Carroll M, et al. (2009). Quantifying the impact of selection bias caused by nonparticipation in a case-control study of mobile phone use. Ann Epidemiol. 19(1):33–41. doi:10.1016/j.annepidem.2008.10.006 PMID:19064187