SCIENTIFIC OPINION

ADOPTED: 20 September 2017

doi: 10.2903/j.efsa.2017.5007

Scientific Opinion of the PPR Panel on the follow-up of the findings of the External Scientific Report 'Literature review of epidemiological studies linking exposure to pesticides and health effects'

EFSA Panel on Plant Protection Products and their Residues (PPR), Colin Ockleford, Paulien Adriaanse, Philippe Berny, Theodorus Brock, Sabine Duquesne, Sandro Grilli, Susanne Hougaard, Michael Klein, Thomas Kuhl, Ryszard Laskowski, Kyriaki Machera, Olavi Pelkonen, Silvia Pieper, Rob Smith, Michael Stemmer, Ingvar Sundh, Ivana Teodorovic, Aaldrik Tiktak, Chris J. Topping, Gerrit Wolterink, Matteo Bottai, Thorhallur Halldorsson, Paul Hamey, Marie-Odile Rambourg, Ioanna Tzoulaki, Daniele Court Marques, Federica Crivellente, Hubert Deluyker and Antonio F. Hernandez-Jerez

Abstract

In 2013, EFSA published a comprehensive systematic review of epidemiological studies published from 2006 to 2012 investigating the association between pesticide exposure and many health outcomes. Despite the considerable amount of epidemiological information available, the quality of much of this evidence was rather low and many limitations likely affect the results so firm conclusions cannot be drawn. Studies that do not meet the 'recognised standards' mentioned in the Regulation (EU) No 1107/2009 are thus not suited for risk assessment. In this Scientific Opinion, the EFSA Panel on Plant Protection Products and their residues (PPR Panel) was requested to assess the methodological limitations of pesticide epidemiology studies and found that poor exposure characterisation primarily defined the major limitation. Frequent use of case–control studies as opposed to prospective studies was considered another limitation. Inadequate definition or deficiencies in health outcomes need to be avoided and reporting of findings could be improved in some cases. The PPR Panel proposed recommendations on how to improve the quality and reliability of pesticide epidemiology studies to overcome these limitations and to facilitate an appropriate use for risk assessment. The Panel recommended the conduct of systematic reviews and meta-analysis, where appropriate, of pesticide observational studies as useful methodology to understand the potential hazards of pesticides, exposure scenarios and methods for assessing exposure, exposure–response characterisation and risk characterisation. Finally, the PPR Panel proposed a methodological approach to integrate and weight multiple lines of evidence, including epidemiological data, for pesticide risk assessment. Biological plausibility can contribute to establishing causation.

Keywords: epidemiology, pesticides, risk assessment, quality assessment, evidence synthesis, lines of evidence, weight-of-evidence

Requestor: European Food Safety Authority

Question number: EFSA-Q-2014-00481

Correspondence: pesticides.ppr@efsa.europa.eu

Acknowledgements: The Panel wishes to thank the following EFSA staff for the support provided to this scientific output: Andrea Terron, Andrea Altieri and Arianna Chiusolo. The Panel and EFSA wishes to acknowledge the following Hearing Experts for their input: (1) David Miller (US-EPA) for sharing the experience of US-EPA and for effect size magnification, (2) Kent Thomas (US-EPA) for the Agricultural Health Study, (3) Marie Christine Lecomte (INSERM), Sylvaine Cordier (INSERM) and Alexis Elbaz (INSERM) for the INSERM Report, (4) Toby Athersuch (Imperial College) for Exposome and Metabolomics, (5) Peter Floyd (Risk & Policy Analysts Ltd), Ruth Bevan (IEH Consulting Ltd), Kate Jones (UK Health & Safety Laboratory) for the Human Biomonitoring data collection from occupational exposure to pesticides. Finally, EFSA would like to thank the Scientific Committee and AMU Unit for the revision of the Opinion and the input provided.

Suggested citation: EFSA PPR Panel (EFSA Panel on Plant Protection Products and their Residues), Ockleford C, Adriaanse P, Berny P, Brock T, Duquesne S, Grilli S, Hougaard S, Klein M, Kuhl T, Laskowski R, Machera K, Pelkonen O, Pieper S, Smith R, Stemmer M, Sundh I, Teodorovic I, Tiktak A, Topping CJ, Wolterink G, Bottai M, Halldorsson T, Hamey P, Rambourg M-O, Tzoulaki I, Court Marques D, Crivellente F, Deluyker H and Hernandez-Jerez AF, 2017. Scientific Opinion of the PPR Panel on the follow-up of the findings of the External Scientific Report 'Literature review of epidemiological studies linking exposure to pesticides and health effects'. EFSA Journal 2017;15(10):5007, 101 pp. https://doi.org/10.2903/j.efsa.2017.5007

ISSN: 1831-4732

The EFSA Journal is a publication of the European Food Safety Authority, an agency of the European Union.

Summary

The European Food Safety Authority (EFSA) asked the Panel on Plant Protection Products and their Residues (PPR Panel) to develop a Scientific Opinion on the follow-up of the findings of the External Scientific Report 'Literature review of epidemiological studies linking exposure to pesticides and health effects' (Ntzani et al., 2013). This report was based on a systematic review and meta-analysis of epidemiological studies published between 2006 and 2012 and summarised the associations found between pesticide exposure and 23 major categories of human health outcomes. Most relevant significant associations were found for liver cancer, breast cancer, stomach cancer, amyotrophic lateral sclerosis, asthma, type II diabetes, childhood leukaemia and Parkinson's disease. While the inherent weaknesses of the epidemiological studies assessed do not allow firm conclusions to be drawn on causal relationships, the systematic review raised a concern about the suitability of regulatory studies to inform on specific and complex human health outcomes.

The PPR Panel developed a Scientific Opinion to address the methodological limitations affecting the quality of epidemiological studies on pesticides. This Scientific Opinion is intended only to assist the peer review process during the renewal of pesticides under Regulation (EC) 1107/2009 where the evaluation of epidemiological studies, along with clinical cases and poisoning incidents following any kind of human exposure, if available, is a data requirement. Epidemiological data concerning exposures to pesticides in Europe will not be available before first approval of an active substance and so will not be expected to contribute to a draft assessment report (DAR). However, there is the possibility that earlier prior approval has been granted for use of an active substance in another jurisdiction and epidemiological data from that area may be considered relevant. Regulation (EC) No 1107/2009 requires a search of the scientific peer-reviewed open literature, which includes existing epidemiological studies. This type of data is more suited for the renewal process of active substances, also in compliance with Regulation (EC) 1141/2010 which indicates that 'The dossiers submitted for renewal should include new data relevant to the active substance and new risk assessments'.

In this Opinion, the PPR Panel proposed a methodological approach specific for pesticide active substances to make appropriate use of epidemiological data for risk assessment purposes, and proposed recommendations on how to improve the quality and reliability of epidemiological studies on pesticides. In addition, the PPR Panel discussed and proposed a methodology for the integration of epidemiological evidence with data from experimental toxicology as both lines of evidence can complement each other for an improved pesticide risk assessment process.

First, the opinion introduces the basic elements of observational epidemiological studies[1] and contrasts them with interventional studies which are considered to provide the most reliable evidence in epidemiological research as the conditions for causal inference are usually met. The major observational study designs are described together with the importance of a detailed description of pesticide exposure, the use of validated health outcomes and appropriate statistical analysis to model exposure–health relationships. The external and internal study validity is also addressed to account for the role of chance in the results and to ascertain whether factors other than exposure can distort the associations found. Several types of human data can contribute to the risk assessment process of pesticides, particularly to support hazard identification. Besides formal epidemiological studies, other sources of human data such as case series, disease registries, poison control centre information, occupational health surveillance data and post-marketing surveillance programmes, can provide useful information for hazard identification, particularly in the context of acute, specific health effects.

However, many of the existing epidemiological studies on pesticides exposure and health effects suffer from a range of methodological limitations or deficiencies (Terms of Reference (ToR) 1). The Panel notes that the complexity of studying associations between exposure to pesticides and health outcomes in observational settings among humans is more challenging than in many other disciplines of epidemiology. This complexity lies in some specific characteristics in the field of pesticide epidemiology such as the large number of active substances in the market (around 480 approved for use in the European Union (EU)), the difficulties to measure exposure, and the frequent lack of quantitative (and qualitative) data on exposure to individual pesticides. The systematic appraisal of epidemiological evidence carried out in an EFSA external scientific report (Ntzani et al., 2013) identified a number of methodological limitations. Poor exposure characterisation primarily defines the major limitation of most existing studies because of the lack of direct and detailed exposure assessment to specific pesticides (e.g. use of generic pesticide definitions). Frequent use of case–control studies as

[1] This Opinion deals only with observational studies (also called epidemiological studies) and vigilance data. In contrast, interventional studies (also called experimental studies, such as randomised clinical trials) are outside the scope of this Opinion.

opposed to prospective studies is also a limitation. Inadequate definition or deficiencies in health outcomes, deficiencies in statistical analysis and poor quality reporting of research findings were identified as other limitations of some pesticide epidemiological studies. These limitations are to some extent responsible for heterogeneity or inconsistency of data that challenge drawing robust conclusions on causality. Given the small effect sizes for most of the outcomes addressed by Ntzani et al. (2013), the contribution of bias in the study design can play a role.

The PPR Panel also provides a number of refinements (ToR 2) and recommendations (ToR 3) to improve future pesticide epidemiological studies that will benefit the risk assessment. The quality and relevance of epidemiological research can be enhanced by (a) an adequate assessment of exposure, preferentially by using personal exposure monitoring or biomarker concentrations of specific pesticides (or combination of pesticides) at an individual level, reported in a way that minimises misclassification of exposure and allows for dose–response assessment; (b) a sufficiently valid and reliable outcome assessment (well defined clinical entities or validated surrogates); (c) adequately accounting for potentially confounding variables (including other known exposures affecting the outcomes); (d) conducting and reporting subgroup analysis (e.g. stratification by gender, age, etc.). A number of reporting guidelines and checklists developed specifically for studies on environmental epidemiology are of interest for epidemiological studies assessing pesticide exposures. This is the case for extensions of the modified STROBE (STrengthening the Reporting of OBservational studies in Epidemiology) criteria, among others, which includes recommendations on what should be included in an accurate and complete report of an observational study.

Exposure assessment can be improved at the individual level (direct and detailed exposure assessment to specific pesticides in order to provide a reliable dosimeter for the pesticide of concern that can be supplemented with other direct measures such as biomonitoring). Besides, exposure can be assessed at population level by using registered data that can then be linked to electronic health records. This will provide studies with unprecedented sample size and information on exposure and subsequent disease. Geographical information systems (GIS) and small area studies might also serve as an additional way to provide estimates of residential exposures. These more generic exposure assessments have the potential to identify general risk factors and may be important both informing overall regulatory policies, and for identification of matters for further epidemiological research. The development of -omic technologies also presents intriguing possibilities for improving exposure assessment through measurement of a wide range of molecules, from xenobiotics and metabolites in biological matrices (metabolomics) to complexes with DNA and proteins (adductomics). Omics have the potential to measure profiles or signatures of the biological response to the cumulative exposure to complex chemical mixtures and allows a better understanding of biological pathways. Health outcomes can be refined by using validated biomarkers of effect, that is, a quantifiable biochemical, physiological or any other change that, is related to level of exposure, is associated with a health impairment and also helps to understand a mechanistic pathway of the development of a disease.

The incorporation of epidemiological studies into regulatory risk assessment (ToR 4) represents a major challenge for scientists, risk assessors and risk managers. The findings of the different epidemiological studies can be used to assess associations between potential health hazards and adverse health effects, thus contributing to the risk assessment process. Nevertheless, and despite the large amount of available data on associations between pesticide exposure and human health outcomes, the impact of such studies in regulatory risk assessment is still limited. Human data can be used for many stages of risk assessment; however, a single (not replicated) epidemiological study, in the absence of other studies on the same pesticide active substance, should not be used for hazard characterisation unless it is of high quality and meets the 'recognised standards' mentioned in the Regulation (EU) No 1107/2009. As these 'recognised standards' are not detailed in the Regulation, a number of recommendations should be considered for optimal design and reporting of epidemiological studies to support regulatory assessment of pesticides. Although further specific guidance will be helpful, this is beyond the ToR of this Opinion. Evidence synthesis techniques, such as systematic reviews and meta-analysis (where appropriate) offer a useful approach. While these tools allow generation of summary data, increased statistical power and precision of risk estimates by combining the results of all individual studies meeting the selection criteria, they cannot overcome methodological flaws or bias of individual studies. Systematic reviews and meta-analysis of observational studies have the capacity of large impact on risk assessment as these tools provide information that strengthens the understanding of the potential hazards of pesticides, exposure scenarios and methods for assessing exposure, exposure–response characterisation and risk characterisation. Although systematic reviews

are also considered a potential tool for answering toxicological questions, their methodology would need to be adapted to the different lines of evidence.

Study evaluation should be performed within a best evidence synthesis framework as it provides an indication on the nature of the potential biases each specific study may have and an assessment of overall confidence in the epidemiological database. This Opinion reports the study quality parameters to be evaluated in single epidemiological studies and the associated weight (low, medium and high) for each parameter. Three basic categories are proposed as a first tier to organise human data with respect to risk of bias and quality: (a) low risk of bias and high/medium reliability; (b) medium risk of bias and medium reliability; (c) high risk of bias and low reliability because of serious methodological limitations or flaws that reduce the validity of results or make them largely uninterpretable for a potential causal association. These categories are intended to parallel the reliability and relevance rating of each stream of evidence according to the EFSA peer review of active substances: acceptable, supplementary and unacceptable. Risk assessment should not be based on results of epidemiological studies that do not meet well-defined data quality standards in order to meet the 'recognised standards' mentioned in the Regulation (EU) No 1107/2009.

Epidemiological studies provide complementary data that can be integrated together with data from *in vivo* laboratory animal studies, mechanistic *in vitro* models and ultimately *in silico* technology for pesticide risk assessment (ToR 4). The combination of all these lines of evidence can contribute to a Weight-of-Evidence (WoE) analysis in the characterisation of human health risks with the aim of improving decision-making. Although the different sets of data can be complementary and confirmatory, and thus serve to strengthen the confidence of one line of evidence on another, they may individually be insufficient and pose challenges for characterising properly human health risks. Hence, all four lines of evidence (epidemiology, animal, *in vitro*, *in silico*) make a powerful combination, particularly for chronic health effects of pesticides, which may take decades to be clinically manifested in an exposed human population.

The first consideration is how well the health outcome under consideration is covered by existing toxicological and epidemiological studies on pesticides. When both types of studies are available for a given outcome/endpoint, both should be assessed for strengths and weaknesses before being used for risk assessment. Once the reliability of available human evidence (observational epidemiology and vigilance data), experimental evidence (animal and *in vitro* data) and non-testing data (*in silico* studies) has been evaluated, the next step involves weighting these sources of data. This opinion proposed an integrated approach where all lines of evidence are considered in an overall WoE framework to better support the risk assessment. This framework relies on a number of principles highlighting when one line should take precedence over another. The concordance or discordance between human and experimental data should be assessed in order to determine which data set should be given precedence. Although the totality of evidence should be assessed, the more reliable data should be given more weight, regardless of whether the data comes from human or experimental studies. The more challenging situation is when study results are not concordant. In such cases, the reasons for the difference should be considered and efforts should be made to develop a better understanding of the biological basis for the contradiction.

Human data on pesticides can help verify the validity of estimations made based on extrapolation from the full toxicological database regarding target organs, dose–response relationships and the reversibility of toxic effects, and to provide reassurance on the extrapolation process without direct effects on the definition of reference values. Thus, pesticide epidemiological data can form part of the overall WoE of available data using modified Bradford Hill criteria as an organisational tool to increase the likelihood of an underlying causal relationship.

Table of contents

1. Introduction

1.1. Regulatory data requirements regarding human health in pesticide risk assessment

Regulatory authorities in developed countries conduct a formal human risk assessment for each registered pesticide based on mandated toxicological studies, done according to specific study protocols, and estimates of likely human exposure.

In the European Union (EU), the procedure for the placing of plant protection products (PPP) on the market is laid down by Commission Regulation No 1107/2009[2]. Commission Regulations No 283/2013[3] and 284/2013[4] set the data requirements for the evaluation and re-evaluation of active substances and their formulations.

The data requirements regarding mammalian toxicity of the active substance are described in part A of Commission Regulation (EU) No 283/2013 for chemical active substances and in part B for microorganisms including viruses. With regard to the requirements for pesticide active substances, reference to the use of human data may be found in different chapters of Section 5 related to different end-points. For instance, data on toxicokinetics and metabolism that include *in vitro* metabolism studies on human material (microsomes or intact cell systems) belong to Chapter 5.1 that deals with studies of absorption, distribution, metabolism and excretion in mammals; *in vitro* genotoxicity studies performed on human material are described in Chapter 5.4 on genotoxicity testing and specific studies such as acetylcholinesterase inhibition in human volunteers are found in Chapter 5.7 on neurotoxicity studies. Chapter 5.8 refers to supplementary studies on the active substance, and some specific studies, such as pharmacological or immunological investigations.

Although the process of pesticide evaluation is mainly based on experimental studies, human data could add relevant information to that process. The requirements relating to human data are mainly found in Chapter 5.9 'Medical data' of Regulation (EU) No 283/2013. It includes medical reports following accidental, occupational exposure or incidents of intentional self-poisoning as well as monitoring studies such as on surveillance of manufacturing plant personnel and others. The information may be generated and reported through official reports from national poison control centres as well as epidemiological studies published in the open literature. The Regulation requires that 'relevant' information on the effects of human exposure, where available, shall be used to confirm the validity of extrapolations regarding exposure and conclusions with respect to target organs, dose–response relationships, and the reversibility of adverse effects.

Regulation (EU) No 1107/2009 equally states that, 'where available, and supported with data on levels and duration of exposure, and conducted in accordance with recognised standards, epidemiological studies are of particular value and must be submitted'. However, it is clear that there is no obligation for the petitioners to conduct epidemiological studies specific for the active substance undergoing the approval or renewal process. Rather, according to Regulation (EC) No 1107/2009, applicants submitting dossiers for approval of active substances shall provide 'scientific peer-reviewed public available literature [. . .]. This should be on the active substance and its relevant metabolites dealing with side-effects on health [. . .] and published within the last ten years before the date of submission of the dossier'.

In particular, epidemiological studies on pesticides should be retrieved from the literature according to the EFSA Guidance entitled 'Submission of scientific-peer reviewed open literature for the approval of pesticide active substances under Regulation (EC) No 1107/2009' (EFSA, 2011a), which follows the principles of the Guidance 'Application of systematic review methodology to food and feed safety assessments to support decision-making' (EFSA, 2010a). As indicated in the EFSA Guidance, 'the process of identifying and selecting scientific peer-reviewed open literature for active substances, their metabolites, or plant protection products' is based on a literature review which is systematic in the approach.

The submission of epidemiological studies and more generally of human data by the applicants in Europe has especially previously sometimes been incomplete and/or has not been performed in

[2] Regulation (EC) No 1107/2009 of the European Parliament and of the Council of 21 October 2009 concerning the placing of plant protection products on the market and repealing Council Directives 79/117/EEC and 91/414/EEC. OJ L 309, 24.11.2009, p. 1–50.

[3] Commission Regulation (EU) No 283/2013, of 1 March 2013, setting out the data requirements for active substances, in accordance with Regulation (EC) No 1107/2009 of the European Parliament of the Council concerning the placing of plant protection products on the market. OJ L 93, 3.4.2013, p. 1–84.

[4] Commission Regulation (EU) No 284/2013 of 1 March 2013 setting out the data requirements for plant protection products, in accordance with Regulation (EC) No 1107/2009 of the European Parliament and of the Council concerning the placing of plant protection products on the market. OJ L 93, 3.4.2013, p. 85–152.

compliance with current EFSA Guidance (EFSA, 2011a). This is probably owing to the fact that a mandatory requirement to perform an (epidemiological) literature search according to specific EFSA Guidance is relatively recent, e.g. introduced for AIR-3 substances (Regulation AIR-3: Reg. (EU) No 844/2012; Guidance Document SANCO/2012/11251 – rev.4).

The integration of epidemiological data with toxicological findings in the peer review process of pesticides in the EU should be encouraged but is still lacking. A recent and controversial example is the one related to the evaluation of glyphosate in which significant efforts were made to include epidemiological studies in the risk assessment, but the conclusion was that these studies provided very limited evidence of an association between glyphosate and health outcomes.

In the case of the peer review of 2,4-D, most of epidemiological data were not used in the risk assessment because it was critical to know the impurity profile of the active substance and this information was not available in the publications (as happens frequently in epidemiological studies). In conclusion, within the European regulatory system there is no example of a pesticide active substance approval being influenced by epidemiological data.

Now that a literature search including epidemiological studies is mandatory and guidance is in place (EFSA, 2011a), a more consistent approach can facilitate risk assessment. However, no framework has been established on how to assess such epidemiological information in the regulatory process. In particular, none of the classical criteria used for the evaluation of these studies is included in the current regulatory framework (e.g. study design, use of odd ratios and relative risks, potential confounders, multiple comparisons, assessment of causality). It follows that specific criteria or guidance for the appropriate use of epidemiological findings in the process of writing and peer reviewing Draft Assessment Reports (DARs) or Renewal Assessment Reports (RAR) is warranted. The EFSA Stakeholder Workshop (EFSA, 2015a) anticipated that the availability of more robust and methodologically sound studies presenting accurate information on exposure would bolster the regulation of pesticides in the EU.

Another potential challenge is synchronisation between the process of renewal of active substances and the output of epidemiological studies. Indeed, the planning, conduct, and analysis of epidemiological studies often require a substantial amount of time, especially where interpretation of data is complex.

1.2. Background and Terms of Reference as provided by the requestor

In 2013, the European Food Safety Authority (EFSA) published an External scientific report 'Literature review on epidemiological studies linking exposure to pesticides and health effects' carried out by the University of Ioannina Medical School (Ntzani et al., 2013). The report is based on a systematic review of epidemiological studies published between 2006 and 2012 and summarises the association between pesticide exposure and any health outcome examined (23 major categories of human health outcomes). In particular, a statistically significant association was observed through fixed and random effect meta-analyses between pesticide exposure and the following health outcomes: liver cancer, breast cancer, stomach cancer, amyotrophic lateral sclerosis, asthma, type II diabetes, childhood leukaemia and Parkinson's disease.

Despite the large number of research articles and analyses (> 6,000) available, the authors of the report could not draw any firm conclusions for the majority of the health outcomes. This observation is in line with previous studies assessing the association between the use of pesticides and the occurrence of human health adverse effects which all acknowledge that such epidemiological studies suffer from a number of limitations and large heterogeneity of data. The authors especially noted that broad pesticides definitions in the epidemiological studies limited the value of the results of meta-analyses. Also, the scope of the report did not allow the in-depth associations between pesticide exposure and specific health outcomes. Nonetheless, the report highlights a number of health outcomes where further research is needed to draw firmer conclusions regarding their possible association with pesticide exposures.

Nevertheless, the outcomes of the External scientific report are in line with other similar studies published in Europe,[5,6] and raise a number of questions and concerns, with regard to pesticide exposure and the associations with human health outcomes. Furthermore, the results of the report

[5] France: INSERM report 2013: Pesticides – effets sur la santé.

[6] UK: COT report 2011: Statement on a systematic review of the epidemiological literature on para-occupational exposure to pesticides and health outcomes other than cancer, and COT report 2006: Joint Statement on Royal Commission on Environmental Pollution report on crop spraying and the health of residents and bystanders.

open the way for discussion on how to integrate results from epidemiological studies into pesticide risk assessments. This is particularly important for the peer-review team at EFSA dealing with the evaluation of approval of plant protection products for which the peer-review needs to evaluate epidemiological findings according to EU Regulation No 283/2013. The regulation states that applicants must submit 'relevant' epidemiological studies, where available.

For the Scientific Opinion, the PPR Panel will discuss the associations between pesticide exposure and human health effects observed in the External scientific report (Ntzani et al., 2013) and how these findings could be interpreted in a regulatory pesticide risk assessment context. Hence, the PPR Panel will systematically assess the epidemiological studies collected in the report by addressing major data gaps and limitations of the studies and provide related recommendations.

The PPR Panel will specifically:

1) collect and review all sources of gaps and limitations, based on (but not necessarily limited to) those identified in the External scientific report in regard to the quality and relevance of the available epidemiological studies.
2) based on the gaps and limitations identified in point 1, propose potential refinements for future epidemiological studies to increase the quality, relevance and reliability of the findings and how they may impact pesticide risk assessment. This may include study design, exposure assessment, data quality and access, diagnostic classification of health outcomes, and statistical analysis.
3) identify areas in which information and/or criteria are insufficient or lacking and propose recommendations for how to conduct pesticide epidemiological studies in order to improve and optimise the application in risk assessment. These recommendations should include harmonisation of exposure assessment (including use of biomonitoring data), vulnerable population subgroups and/or health outcomes of interest (at biochemical, functional, morphological and clinical level) based on the gaps and limitations identified in point 1.
4) discuss how to make appropriate use of epidemiological findings in risk assessment of pesticides during the peer review process of draft assessment reports, e.g. weight-of-evidence (WoE) as well as integrating the epidemiological information with data from experimental toxicology, adverse outcome pathways (AOP), mechanism of actions, etc.

The PRAS Unit will consult the Scientific Committee on the consensual approach to EFSA's overarching scientific areas,[7] including the integration of epidemiological studies in risk assessment.

1.3. Interpretation of the Terms of Reference

In the Terms of Reference (ToR), EFSA requested the PPR Panel to write a scientific Opinion on the follow up of the results from the External Scientific Report on a systematic review of epidemiological studies published between 2006 and 2012 linking exposure to pesticides and human health effects (Ntzani et al., 2013). According to EU Regulation No 283/2013, the integration of epidemiological data into pesticide risk assessment is important for the peer review process of DAR and RAR of active substances for EU approval and their intended use as plant protection products.

In its interpretation of the terms of reference, the PPR Panel will then develop a Scientific Opinion to address the methodological limitations identified in epidemiological studies on pesticides and to make recommendations to the sponsors of such studies on how to improve them in order to facilitate their use for regulatory pesticide risk assessment, particularly for substances in the post-approval period. The PPR Panel notes that experimental toxicology studies also present limitations related to their methodology and quality of reporting; however, the assessment of these limitations is beyond the ToR of this Opinion.

This Scientific Opinion is intended to assist the peer review process during the renewal of pesticides under Regulation 1107/2009 where the evaluation of epidemiological studies, along with clinical cases and poisoning incidents following any kind of human exposure, if available, represent a data requirement. Epidemiological data concerning exposures to pesticides in Europe will not be available before first approval of an active substance (with the exception of incidents produced during the manufacturing process, which are expected to be very unlikely) and so will not be expected to contribute to a DAR. However, there is the possibility that earlier prior approval has been granted for use of an active substance in another jurisdiction and epidemiological data from that area may be considered relevant. Regulation (EC) No 1107/2009 requires a search of the scientific peer-reviewed

[7] According to article 28 of Regulation (EC) No 178/2002.

open literature, where it is expected to retrieve existing epidemiological studies. It is therefore recognised that epidemiological studies are more suitable for the renewal process of active substances, also in compliance with the provision of the EC regulation 1141/2010 indicating that 'The dossiers submitted for renewal should include new data relevant to the active substance and new risk assessments to reflect any changes in data requirements and any changes in scientific or technical knowledge since the active substance was first included in Annex I to Directive 91/414/EEC'.

The PPR Panel will specifically address the following topics:

1) Review inherent weaknesses affecting the quality of epidemiological studies (including gaps and limitations of the available pesticide epidemiological studies) and their relevance in the context of regulatory pesticide risk assessment. How can these weaknesses be addressed?
2) What are potential contributions of epidemiological studies that complement classical toxicological studies conducted in laboratory animal species in the area of pesticide risk assessment?
3) Discuss and propose a methodological approach specific for pesticide active substances on how to make appropriate use of epidemiological studies, focusing on how to improve the gaps and limitations identified.
4) Propose refinements to practice and recommendations for better use of the available epidemiological evidence for risk assessment purposes. Discuss and propose a methodology for the integration of epidemiological information with data from experimental toxicology.

This Scientific Opinion, particularly Section 2–4, is not intended to address the bases of epidemiology as a science. Those readers willing to deepen into specific aspects of this science are encouraged to read general textbook of epidemiology (e.g. Rothman et al., 2008).

It should be taken into account that this Opinion is focussed only on pesticide epidemiology studies in the EU regulatory context and not from a general scientific perspective. Therefore, the actual limitations and weaknesses of experimental toxicology studies will not be addressed herein.

1.4. Additional information

In order to fully address topics 1–4 above (Section 1.3), attention has been paid to a number of relevant reviews of epidemiological studies and the experience of other National and International bodies with knowledge of epidemiology in general and in applying epidemiology to pesticide risk assessment specifically. Detailed attention has been given to these studies in Annex A and drawn from the experience of the authors that have contributed constructively to understanding in this area. Also Annex A records published information that has been criticised for its lack of rigour showing how unhelpful some published studies may be. The lessons learned from such good (and less-good) practice have been incorporated into the main text by cross-referring to Annex A. In this way, this Scientific Opinion has the aim of clearly distilling and effectively communicating the arguments in the main text without overwhelming the reader with all the supporting data which is nevertheless accessible.

In addition, Annex B contains a summary of the main findings of a project that EFSA outsourced in 2015 to further investigate the role of human biological monitoring (HBM) in occupational health and safety strategies as a tool for refined exposure assessment in epidemiological studies and to contribute to the evaluation of potential health risks from occupational exposure to pesticides (Bevan et al., 2017).

2. General framework of epidemiological studies on pesticides

This section introduces the basic elements of epidemiological studies on pesticides and contrasts them with other types of studies. For more details general textbook on epidemiology are recommended (Rothman et al., 2008; Thomas, 2009).

2.1. Study design

Epidemiology studies the distribution and determinants of health outcomes in human or other target species populations, to ascertain how, when and where diseases occur. This can be done through observational studies and intervention studies (i.e. clinical trials),[8] which compare study

[8] In this opinion, 'human data' includes observational studies, also called epidemiological studies, where the researcher is observing natural relationships between factors and health outcomes without acting upon study participants. Vigilance data also fall under this concept. In contrast, intervention studies (also referred to as experimental studies) are outside the scope of this Opinion, and their main feature is that the researcher intercedes as part of the study design.

groups subject to differing exposure to a potential risk factor. Both types of studies are carried out in a natural setting, which is a less controlled environment than laboratories.

Information on cases of disease occurring in a natural setting can also be systematically recorded in the form of case reports or case series of exposed individuals only. Although case series/reports do not compare study groups according to differing exposure, they may provide useful information, particularly on acute effects following high exposures, which makes them potentially relevant for hazard identification.

In randomised clinical trials, the exposure of interest is randomly allocated to subjects and, whenever possible, these subjects are blinded to their treatment, thereby eliminating potential bias due to their knowledge about their exposure to a particular treatment. This is why they are called intervention studies. Observational epidemiological studies differ from clinical intervention studies in that the exposure of interest is not randomly assigned to the subjects enrolled and participants are often not blinded to their exposure. This is why they are called observational. As a result, randomised clinical trials rank higher in terms of design as they provide unbiased estimates of average treatment effects.

The lack of random assignment of exposure in observational studies represents a key challenge, as other risk factors that are associated with the occurrence of disease may be unevenly distributed between those exposed and non-exposed. This means that known confounders need to be measured and accounted for. However, there is always the possibility that unknown or unmeasured confounders are left unaccounted for, although unknown confounders cannot be addressed. Furthermore, the fact that study participants are often unaware of their current or past exposure or may not recall these accurately in observational studies (e.g. second-hand smoke, dietary intake or occupational hazards) may result in biased estimates of exposure if it is based on self-report. As an example, it is not unlikely that when cancer cases and controls are asked whether they have previously been exposed to a pesticide the cancer cases may report their exposure differently from controls, even in cases where the past exposures did not differ between the two groups.

Traditionally, designs of observational epidemiological studies are classified as either ecological, cross-sectional, case–control or cohort studies. This approach is based on the quality of exposure assessment and the ability to assess directionality from exposure to outcome. These differences largely determine the quality of the study (Rothman and Greenland, 1998; Pearce, 2012).

- **Ecological studies** are observational studies where either exposure, outcome or both are measured on a group but not at individual level and the correlation between the two is then examined. Most often, exposure is measured on a group level while the use of health registries often allows for extraction of health outcomes on an individual level (cancer, mortality). These studies are often used when direct exposure assessment is difficult to achieve and in cases where large contrast in exposures are needed (comparing levels between different countries or occupations). Given the lack of exposure and/or outcome on an individual level, these studies are useful for hypothesis generation but results generally need to be followed up using more rigorous design in either humans or use of experimental animals.
- In **cross-sectional studies,** exposure and health status are assessed at the same time, and prevalence rates (or incidence over a limited recent time) in groups varying in exposure are compared. In such studies, the temporal relationship between exposure and disease cannot be established since the current exposure may not be the relevant time window that leads to development of the disease. The inclusion of prevalent cases is a major drawback of (most) cross-sectional studies, particularly for chronic long-term diseases. Cross-sectional studies may nevertheless be useful for risk assessment if exposure and effect occur more or less simultaneously or if exposure does not change over time.
- **Case–control studies** examine the association between estimates of past exposures among individuals that already have been diagnosed with the outcome of interest (e.g. cases) to a control group of subjects from the same population without such outcome. In population-based incident case–control studies, cases are obtained from a well-defined population, with controls selected from members of the population who are disease free at the time a case is incident. The advantages of case–control studies are that they require less sample sizes, time and resources compared to prospective studies and often they are the only viable option when studying rare outcomes such as some types of cancer. In case–control studies, past exposure is most often not assessed based on 'direct' measurement but rather through less certain measurements such as a recall captured through interviewer or self-administered

questionnaires or proxies such as job descriptions titles or task histories. Although case–control studies may allow for proper exposure assessment, these studies are prone to recall-bias when estimating exposure. Other challenges include the selection of appropriate controls; as well as the need for appropriate confounder control.

- In **cohort studies**, the population under investigation consists of individuals who are at risk of developing a specific disease or health outcome at some point in the future. At baseline and at later follow-ups (prospective cohort studies) relevant exposures, confounding factors and health outcomes are assessed. After an appropriate follow-up period, the frequency of occurrence of the disease is compared among those differently exposed to the previously assessed risk factor of interest. Cohort studies are therefore by design prospective as the assessment of exposure to the risk factor and covariates of interest are measured before the health outcome has occurred. Thus, they can provide better evidence for causal associations compared to the other designs mentioned above. In some cases, cohort studies may be based on estimates of past exposure. Such retrospective exposure assessment is less precise than direct measure and prone to recall bias. As a result, the quality of evidence from cohort studies varies according to the actual method used to assess exposure and the level of detail by which information on covariates were collected. Cohort studies are particularly useful for the study of relatively common outcomes. If sufficiently powered in terms of size, they can also be used to appropriately address relatively rare exposures and health outcomes. Prospective cohort studies are also essential to study different critical exposure windows. An example of this is longitudinal birth cohorts that follow children at regular intervals until adult age. Cohort studies may require a long observation period when outcomes have a long latency prior to onset of disease. Thus, such studies are both complex and expensive to conduct and are prone to loss of follow-up.

2.2. Population and sample size

A key strength of epidemiological studies is that they study diseases in the very population about which conclusions are to be drawn, rather than a proxy species. However, only rarely will it be possible to study the whole population. Instead, a sample will be drawn from the reference population for the purpose of the study. As a result, the observed effect size in the study population may differ from that in the population if the former does not accurately reflect the latter. However, observations made in a non-representative sample may still be valid within that sample but care should then be made when extrapolating findings to the general population.

Having decided how to select individuals for the study, it is also necessary to decide how many participants should minimally be enrolled. The sample size of a study should be large enough to warrant sufficient statistical power. The standard power (also called sensitivity) is 80%, which means the ability of a study to detect an effect of a given magnitude when that effect actually exists in the target population; in other words, there is 80% probability of drawing the right conclusion from the results of the analyses and a corresponding probability of 20% or drawing the wrong conclusion and missing a true effect. Power analysis is often used to calculate the minimum sample size required to likely detect an effect of a given size. Small samples are likely to constitute an unrepresentative sample. The statistical power is also closely related to risk inflation, which needs to be given special attention when interpreting statistically significant results from small or underpowered studies (see Annex D).

Epidemiological studies, like toxicological studies in laboratory animals, are often designed to examine multiple endpoints unlike clinical trials that are designed and conducted to test one single hypothesis, e.g. efficacy of a medical treatment. To put this in context, for laboratory animal toxicology test protocols, OECD guidance for pesticides may prescribe a minimum number of animals to be enrolled in each treatment group. This does not guarantee adequate power for any of the multitude of other endpoints being tested in the same study. It is thus important to appropriately consider the power of a study when conducting both epidemiology and laboratory studies.

2.3. Exposure

The quality of the exposure measurements influences the ability of a study to correctly ascertain the causal relationship between the (dose of) exposure and a given adverse health outcome.

In toxicological studies in laboratory animals, the 'treatment regime' i.e. dose, frequency, duration and route are well defined beforehand and its implementation can be verified. This often allows

expression of exposure in terms of external dose administered daily via oral route for example in a 90-day study, by multiplying the amount of feed ingested every day by a study animal with the intended (and verified) concentration of the chemical present in the feed. Also, in the future, the internal exposure has to be determined in the pivotal studies.

In the case of pesticides, estimating exposure in a human observational setting is difficult as the dose, its frequency and duration over time and the route of exposure are not controlled and not even well known.

Measuring the intensity, frequency and duration of exposure is often necessary for investigating meaningful associations. Exposure may involve a high dose over a relatively short period of time, or a low-level prolonged dose over a period from weeks to years. While the effects of acute, high-dose pesticide exposure may appear within hours or days, the effects of chronic, low-dose exposures may not appear until years later. Also, a disease may require a minimal level of exposure but increase in probability with longer exposure.

There may be differences in absorption and metabolism via different routes (dermal, inhalation and oral). While dermal or inhalation are often the routes exposure occurs in occupational settings, ingestion (food, water) may be the major route of pesticide exposure for the general population. Pharmacokinetic differences among individuals may result in differing systemic or tissue/organ doses even where the absorbed external doses may appear similar.

2.4. Health outcomes

The term health outcome refers to a disease state, event, behaviour or condition associated with health that is under investigation. Health outcomes are those clinical events (usually represented as diagnosis codes, i.e. International Classification of Diseases (ICD) 10) or outcomes (i.e. death) that are the focus of the research. Use of health outcomes requires a well-defined case definition, a system to report and record the cases and a measure to express the frequency of these events.

A well-defined case definition is necessary to ensure that cases are consistently diagnosed, regardless of where, when and by whom they were identified and thus avoid misclassification. A case definition involves a standard set of criteria, which can be a combination of clinical symptoms/signs, sometimes supplemented by confirmatory diagnostic tests with their known sensitivity and specificity. The sensitivity of the whole testing procedure (i.e. the probability that a person with an adverse health condition is truly diagnosed) must be known to estimate the true prevalence or incidence.

The clinical criteria may also involve other characteristics (e.g. age, occupation) that are associated with increased disease risk. At the same time, appropriately measured and defined phenotypes or hard clinical outcomes add validity to the results.

Disease registries contain clinical information of patients on diagnosis, treatment and outcome. These registries periodically update patient information and can thus provide useful data for epidemiological research. Mortality, cancer and other nation-wide health registries generally meet the case-definition requirements and provide (almost) exhaustive data on the incident cases within a population. These health outcomes are recorded and classified in national health statistics databases, which depend on accepted diagnostic criteria that are evolving and differ from one authority to another. This may confound attempts to pool data usefully for societal benefit. Registry data present many opportunities for meaningful analysis, but the degree of data completeness and validity may challenge making appropriate inferences. Also, changes in coding conventions over the lifetime of the database may have an impact on retrospective database research.

Although the disease status is typically expressed as a dichotomous variable, it may also be measured as an ordinal variable (e.g. severe, moderate, mild or no disease) or as a quantitative variable for example by measuring molecular biomarkers of toxic response in target organs or physiological measures such as blood pressure or serum concentration of lipids or specific proteins.

The completeness of the data capture and its consistency are key contributors to the reliability of the study. Harmonisation of diagnostic criteria, data storage and utility would bring benefits to the quality of epidemiological studies.

A surrogate endpoint is used as substitute for a well-defined disease endpoint, an outcome measure, commonly a laboratory measurement (biomarker of response). These measures are considered to be on the causal pathway for the clinical outcome. In contrast to overt clinical disease, such biological markers of health may allow to detect subtle, subclinical toxicodynamic processes. For such outcomes, detailed analytical protocols for quantification should be specified to enable comparison or replication across laboratories. The use of AOPs can highlight differences in case definitions.

Although surrogate outcomes may offer additional information, the suitability of the surrogate outcome examined needs to be carefully assessed. In particular, the validity of surrogate outcomes may represent a major limitation to their use (la Cour et al., 2010). Surrogate endpoints that have not been validated should thus be avoided.

When the health status is captured in other ways, such as from self-completed questionnaires or telephone interviews, from local records (medical or administrative databases) or through clinical examination only, these should be validated to demonstrate that they reflect the underlying case definition.

2.5. Statistical analysis and reporting

Reporting in detail materials, methods and results, and conducting appropriate statistical analyses are key steps to ensure quality of epidemiological studies. Regarding statistical analysis, one can distinguish between descriptive statistics and modelling of exposure–health outcome relationship.

2.5.1. Descriptive statistics

Descriptive statistics aim to summarise the important characteristics of the study groups, such as exposure measures, health outcomes, possible confounding factors and other relevant factors. The descriptive statistics often include frequency tables and measures of central tendency (e.g. means and medians) and dispersion (e.g. variance and interquartile range) of the parameters or variables studied.

2.5.2. Modelling exposure–health outcome relationship

Modelling of the exposure–health relationship aims to assess the possible relationship between the exposure and the health outcome under consideration. In particular, it can evaluate how this relationship may depend on dose and mode of exposure and other possible intervening factors.

Statistical tests determine the probability that the observations found in scientific studies may have occurred as a result of chance. This is done by summarising the results from individual observations and evaluating whether these summary estimates differ significantly between, e.g. exposed and non-exposed groups, after taking into consideration random errors in the data.

For dichotomous outcomes, the statistical analysis compares study groups by assessing whether there is a difference in disease frequency between the exposed and control populations. This is usually done using a relative measure. The relative risk (RR) in cohort studies estimates the relative magnitude of an association between exposure and disease comparing those that are exposed (or those that have a higher exposure level) with those that are not exposed (or those that have a lower exposure level). It indicates the likelihood of developing the disease in the exposed group relative to those who are not (or less) exposed. An odds ratio (OR), generally an outcome measure in case–control and cross-sectional studies, represents the ratio of the odds of exposure between cases and controls (or diseased and non-diseased individuals in a cross-sectional study) and is often the relative measure used in statistical testing. Different levels or doses of exposure can be compared in order to see if there is a dose–response relationship. For continuous outcome measures, mean or median change in the outcome are often examined across different level of exposure; either through analyses of variance or through other parametric statistics.

While the statistical analysis will show that observed differences are significantly different or not significantly different, both merit careful reflection (Greenland et al., 2016).

Interpretation of the absence of statistically significant difference. Failure to reject the null hypothesis does not necessarily mean that no association is present because the study may not have sufficient power to detect it. The power depends on the following factors:

- sample size: with small sample sizes, statistical significance is more difficult to detect, even if true;
- variability in individual response or characteristics, either by chance or by non-random factors: the larger the variability, the more difficult to demonstrate statistical significance;
- effect size or the magnitude of the observed difference between groups: the smaller the size of the effect, the more difficult to demonstrate statistical significance.

Interpretation of statistically significant difference. Statistical significance means that the observed difference is not likely due to chance alone. However, such a result still merits careful consideration.

- Biological relevance. Rejection of the null hypothesis does not necessarily mean that the association is biologically meaningful, nor does it mean that the relationship is causal (Skelly, 2011). The key issue is whether the magnitude of the observed difference (or 'effect size') is large enough to be considered biologically relevant. Thus, an association that is statistically significant may be or may be not biologically relevant and vice versa. While epidemiological results that are statistically significant may be dismissed as 'not biologically relevant', non-statistically significant results are seldom determined to be 'biologically relevant'. Increasingly, researchers and regulators are looking beyond statistical significance for evidence of a 'minimal biologically important difference' for commonly used outcomes measures. Factoring biological significance relevance into study design and power calculations, and reporting results in terms of biological as well as statistical significance will become increasingly important for risk assessment (Skelly, 2011). This is the subject of an EFSA Scientific Committee guidance document outlining generic issues and criteria to be taken into account when considering biological relevance (EFSA Scientific Committee, 2017a); also a framework is being developed to consider biological relevance at three main stages related to the process of dealing with evidence (EFSA Scientific Committee, 2017b).
- Random error. Evaluation of statistical precision involves consideration of random error within the study. Random error is the part of the study that cannot be predicted because that part is attributable to chance. Statistical tests determine the probability that the observations found in scientific studies have occurred as a result of chance. In general, as the number of study participants increases, precision (often expressed as standard error) of the estimate of central tendency (e.g. the mean) is increased and the ability to detect a statistically significant difference, if there is a real difference between study groups, i.e. the study's power, is enhanced. However, there is always a possibility, at least in theory, that the results observed are due to chance only and that no true differences exist between the compared groups (Skelly, 2011). Very often this value is set at 5% (significance level).
- Multiple testing. As mentioned previously when discussing sample size, modelling of the exposure–health relationship is in principle hypothesis-driven, i.e. it is to be stated beforehand in the study objectives what will be tested. However, in reality, epidemiological studies (and toxicological studies in laboratory animals) often explore a number of different health outcomes in relation to the same exposure. If many statistical tests are conducted, some 5% of them will be statistically significant by chance. Such testing of multiple endpoints (hypotheses) increases the risk of false positive results and this can be controlled for by use of Bonferroni, Sidak or Benjamini–Hochberg corrections or other suitable methods. But this is often omitted. Thus, when researchers carry out many statistical tests on the same set of data, they can conclude that there are real differences where in fact there are none. Therefore, it is important to consider large number of statistical results as preliminary indications that require further validation. The EFSA opinion on statistical significance and biological significance notes that the assumptions derived from a statistical analysis should be related to the study design (EFSA, 2011b).
- Effect size magnification. An additional source of bias, albeit one that is lesser known, is that which may result from small sample sizes and the consequent low statistical power. This lesser known type of bias is 'effect size magnification' which can result from low powered studies. While it is generally widely known that small, low-powered studies can result in false negatives since the study power is inadequate to reliably detect a meaningful effect size, it is less well known that these studies can result in inflation of effect sizes if those estimated effects pass a statistical threshold (e.g. the common $p < 0.05$ threshold used to judge statistical significance). This effect –also known as effect size magnification – is a phenomenon by which a 'discovered' association (i.e. one that has passed a given threshold of statistical significance) from a study with suboptimal power to make that discovery will produce an observed effect size that is artificially – and systematically – inflated. This is because smaller, low-powered studies are more likely to be affected by random variation among individuals than larger ones. Mathematically, conditional on a result passing some predetermined threshold of statistical significance, the estimated effect size is a biased estimate of the true effect size, with the magnitude of this bias inversely related to power of the study.
  As an example, if a trial were run thousands of times, there will be a broad distribution of observed effect sizes, with smaller trials systematically producing a wider variation in observed effect sizes than larger trials, but the median of these estimated effect sizes is close to the true

effect size. However, in a small and low powered study, only a small proportion of observed effects will pass any given (high) statistical threshold of significance and these will be only the ones with the greatest of effect sizes. Thus, when these smaller, low powered studies with greater random variation do indeed find a significance-triggered association as a result of passing a given statistical threshold, they are more likely to overestimate the size of that effect. What this means is that research findings of small and significant studies are biased in favour of finding inflated effects. In general, the lower the background (or control or natural) rate, the lower the effect size of interest, and the lower the power of the study, the greater the tendency towards and magnitude of inflated effect sizes.
It is important to note, however, that this phenomenon is only present when a 'pre-screening' for statistical significance is done. The bottom line is that if it is desired to estimate a given quantity such as an OR or RR, 'pre-screening' a series of effect sizes for statistical significance will result in an effect size that is systematically biased away from the null (larger than the true effect size). To the extent that regulators, decision-makers, and others are acting in this way – looking for statistically significant results in what might be considered a sea of comparisons and then using those that cross a given threshold of statistical significance to evaluate and judge the magnitude of the effect – will likely result in an exaggerated sense of the magnitude of the hypothesised association. Additional details and several effect size simulations are provided in Annex D of this document.

Confounding occurs when the relationship between the exposure and disease is to some extent attributable to the effect of another risk factor, i.e. the confounder. There are several traditionally recognised requirements for a risk factor to actually act as a confounder as described by McNamee (2003) and illustrated below. The factor must:

- be a cause of the disease, or a surrogate measure of the cause, in unexposed people; factors satisfying this condition are called 'risk factors';
- be correlated, positively or negatively, with exposure in the study populations independently from the presence of the disease. If the study population is classified into exposed and unexposed groups, this means that the factor has a different distribution (prevalence) in the two groups;
- not be an intermediate step in the causal pathway between the exposure and the disease

Confounding can result in an over- or underestimation of the relationship between exposure and disease and occurs because the effects of the two risk factors have not been separated or 'disentangled'. In fact, if strong enough, confounding can also reverse an apparent association. For instance, because agriculture exposures cover many different exposure categories, farmers are likely to be more highly exposed than the general population to a wide array of risk factors, including biological agents (soil organisms, livestock, farm animals), pollen, dust, sunlight and ozone amongst others, which may act as potential confounding factors.

A number of procedures are available for controlling confounding, both in the design phase of the study or in the analytical phase. For large studies, control in the design phase is often preferable. In the design phase, the epidemiological researcher can limit the study population to individuals that share a characteristic which the researcher wishes to control. This is known as 'restriction' and in fact removes the potential effect of confounding caused by the characteristic which is now eliminated. A second method in the design phase through which the researcher can control confounding is by 'matching'. Here, the researcher matches individuals based on the confounding variable which ensures that this is evenly distributed between the two comparison groups.

Beyond the design phase, at the analysis stage, control for confounding can be done by means of either stratification or statistical modelling. One means of control is by stratification in which the association is measured separately, under each of the confounding variables (e.g. males and females, ethnicity or age group). The separate estimates can be 'brought together' statistically – when appropriate – to produce a common OR, RR or other effect size measure by weighting the estimates measured in each stratum (e.g. using Mantel–Haenszel approaches). This can be done at the cost of reducing the sample size for the analysis. Although relatively easy to perform, there can be difficulties associated with the inability of this stratification to deal with multiple confounders simultaneously. For these situations, control can be achieved through statistical modelling (e.g. multiple logistic regression).

Regardless of the approaches available for control of confounding in the design and analysis phases of the study described above, it is important – prior to any epidemiological studies being initiated in the field – that careful consideration be given to confounders because researchers cannot control for a variable which they have not considered in the design or for which they have not collected data.

Epidemiological studies – published or not – are often criticised for ignoring potential confounders that may possibly either falsely implicate or inappropriately negate a given risk factor. Despite these critiques, rarely is an argument presented on the likely size of the impact of the bias from such possible confounding. It should be emphasised that a confounder must be a relatively strong risk factor for the disease to be strongly associated with the exposure of interest to create a substantial distortion in the risk estimate. It is not sufficient to simply raise the possibility of confounding; one should make a persuasive argument explaining why a risk factor is likely to be a confounder, what its impact might be and how important that impact might be to the interpretation of findings. It is important to consider the magnitude of the association as measured by the RR, OR, risk ratio, regression coefficient, etc. since strong relative risks are unlikely to be due to unmeasured confounding, while weak associations may be due to residual confounding by variables that the investigator did not measure or control in the analysis (US-EPA, 2010b).

Effect modification. Effects of pesticides, and other chemicals, on human health can hardly be expected to be identical across all individuals. For example, the effect that any given active substance might have on adult healthy subjects may not be the same as that it may have on infants, elderly, or pregnant women. Thus, some subsets of the population are more likely to develop a disease when exposed to a chemical because of an increased sensitivity. For this, the term 'vulnerable subpopulation' has been used, which means children, pregnant women, the elderly, individuals with a history of serious illness and other subpopulations identified as being subject to special health risks from exposure to environmental chemicals (i.e. because of genetic polymorphisms of drug-metabolising enzymes, transporters or biological targets). The average effect measures the effect of an exposure averaged over all subpopulations. However, there may be heterogeneity in the strength of an association between various subpopulations. For example, the magnitude of the association between exposure to chemical A and health outcome B may be stronger in children than in healthy adults, and absent in those wearing protective clothing at the time of exposure or in those of different genotype. If heterogeneity is truly present, then any single summary measure of an overall association would be deficient and possibly misleading. The presence of heterogeneity is assessed by testing for the presence of statistically significant interaction between the factor and the effect in the various subpopulations. But, in practice, this requires large sample size.

Investigating the effect in subpopulations defined by relevant factors may advance knowledge on the effect on human health of the risk factor of interest.

2.6. Study validity

When either a statistically significant association or no such significant association between, for example, pesticide exposures and a health outcome is observed, there is a need to also evaluate the validity of a research study, assessing factors that might distort the true association and/or influence its interpretation. These imperfections relate to systematic sources of error that result in a (systematically) incorrect estimate of the association between exposure and disease. In addition, the results from a single study takes on increased validity when it is replicated in independent investigations conducted on other populations of individuals at risk of developing the disease.

Temporal sequence. Any claim of causation must involve the cause preceding in time the presumed effect. Rothman (2002) considered temporality as the only criterion that is truly causal, such that lack of temporality rules out causality. While the temporal sequence of an epidemiological association implies the necessity for the exposure to precede the outcome (effect) in time, measurement of the exposure is not required to precede measurement of the outcome. This requirement is easier met in prospective study designs (i.e. cohort studies), than when exposure is assessed retrospectively (case–control studies) or assessed at the same time than the outcome (cross-sectional studies). However, also in prospective studies, the time sequence for cause and effect and the temporal direction might be difficult to ascertain if a disease developed slowly and initial forms of disease were difficult to measure (Höfler, 2005).

The generalisability of the result from the population under study to a broader population should also be considered for study validity. While the random error discussed previously is considered a precision problem and is affected by sampling variability, **bias** is considered a validity issue. More

specifically, bias issues generally involve methodological imperfections in study design or study analysis that affect whether the correct population parameter is being estimated. The main types of bias include selection bias, information bias (including recall bias and interviewer/observer bias) and confounding. An additional potential source of bias is effect size magnification, which has already been mentioned.

Selection bias concerns a systematic error relating to validity that occurs as a result of the procedures and methods used to select subjects into the study, the way that subjects are lost from the study or otherwise influence continuing study participation.

Typically, such a bias occurs in a case–control study when inclusion (or exclusion) of study subjects on the basis of disease is somehow related to the prior exposure status being studied. One example might be the tendency for initial publicity or media attention to a suspected association between an exposure and a health outcome to result in preferential diagnosis of those that had been exposed compared to those that had not. Selection bias can also occur in cohort studies if the exposed and unexposed groups are not truly comparable as when, for example, those that are lost from the study (loss to follow-up, withdrawn or non-response) are different in status to those who remain. Selection bias can also occur in cross-sectional studies due to selective survival: only those that have survived are included in the study. These types of bias can generally be dealt with by careful design and conduct of a study (see also Sections 4, 6 and 8).

The 'healthy worker effect' (HWE) is a commonly recognised selection bias that illustrates a specific bias that can occur in occupational epidemiology studies: workers tend to be healthier than individuals from the general population overall since they need to be employable in a workforce and can thus often have a more favourable outcome status than a population-based sample obtained from the general population. Such a HWE bias can result in observed associations that are masked or lessened compared to the true effect and thus can lead to the appearance of lower mortality or morbidity rates for workers exposed to chemicals or other deleterious substances.

Information bias concerns a systematic error when there are systematic differences in the way information regarding exposure or the health outcome are obtained from the different study groups that result in incorrect or otherwise erroneous information being obtained or measured with respect to one or more covariates being measured in the study. Information bias results in misclassification which in turn leads to incorrect categorisation with respect to either exposure or disease status and thus the potential for bias in any resulting epidemiological effect size measure such as an OR or RR.

Misclassification of exposure status can result from imprecise, inadequate or incorrect measurements; from a subject's incorrect self-report; or from incorrect coding of exposure data.

Misclassification of disease status can, for example, arise from laboratory error, from detection bias, from incorrect or inconsistent coding of the disease status in the database, or from incorrect recall. Recall bias is a type of information bias that concerns a systematic error when the reporting of disease status is different, depending on the exposure status (or vice versa). Interviewer bias is another kind of information bias that occurs where interviewers are aware of the exposure status of individuals and may probe for answers on disease status differentially – whether intended or not – between exposure groups. This can be a particularly pernicious form of misclassification – at least for case–control studies – since a diseased subject may be more likely to recall an exposure that occurred at an earlier time period than a non-diseased subject. This will lead to a bias away from null value (of no relation between exposure and disease) in any effect measure.

Importantly, such misclassifications as described above can be 'differential' or 'non-differential' and these relate to (i) the degree to which a person that is truly exposed (or diseased) is correctly classified as being truly exposed or diseased and (ii) the degree to which an individual who is truly not exposed (or diseased) is correctly classified in that way. The former is known as 'sensitivity' while the latter is referred to as 'specificity' and both of these play a role in determining the existence and possible direction of bias. Differential misclassification means that misclassification has occurred in a way that depends on the values of other variables, while non-differential misclassification refers to misclassifications that do not depend on the value of other variables.

What is important from an epidemiological perspective is that misclassification biases – either differential or non-differential – depend on the sensitivity and specificity of the study's methods used to categorise such exposures and can have a predictable effect on the direction of bias under certain (limited) conditions: this ability to characterise the direction of the bias based on knowledge of the study methods and analyses can be useful to the regulatory decision-maker since it allows the decision maker to determine whether the epidemiological effect sizes being considered (e.g. OR, RR) are likely underestimates or overestimates of the true effect size. While it is commonly assumed by some that

non-differential misclassification bias produces predictable biases towards the null (and thus systematically under-predicts the effect size), this is not necessarily the case. Also, the sometimes common assumption in epidemiology studies that misclassification is non-differential (which is sometimes also paired with the assumption that non-differential misclassification bias is always towards the null) is not always justified (e.g. see Jurek et al., 2005).

When unmeasured confounders are thought to affect the results, researchers should conduct sensitivity analyses to estimate the range of impacts and the resulting range of adjusted effect measures (US-EPA, 2010b). Quantitative sensitivity (or bias) analyses are, however, not typically conducted in many epidemiological studies, with most researchers instead describing various potential biases qualitatively in the form of a narrative in the discussion section of a paper.

It is often advisable that the epidemiological investigator performs sensitivity analysis to estimate the impact of biases, such as exposure misclassification or selection bias, by known but unmeasured risk factors or to demonstrate the potential effects that a missing or unaccounted for confounder may have on the observed effect sizes (see Lash et al., 2009; Gustafson and McCandless, 2010). Sensitivity analyses should be incorporated in the list of criteria for reviewing epidemiological data for risk assessment purposes.

3. Key limitations of the available epidemiological studies on pesticides

3.1. Limitations identified by the authors of the EFSA external scientific report

The EFSA External scientific report (Ntzani et al., 2013; summarised in Annex A) identified a plethora of epidemiological studies which investigate diverse health outcomes. In an effort to systematically appraise the epidemiological evidence, a number of methodological limitations were highlighted. In the presence of these limitations, robust conclusions could not be drawn, but outcomes for which supportive evidence from epidemiology existed were highlighted for future investigation. The main limitations identified included (Ntzani et al., 2013):

- Lack of prospective studies and frequent use of study designs that are prone to bias (case–control and cross-sectional studies). In addition, many of the studies assessed appeared to be insufficiently powered.
- Lack of detailed exposure assessment, at least compared to many other fields within epidemiology. The information on specific pesticide exposure and co-exposures was often lacking, and appropriate biomarkers were seldom used. Instead, many studies relied on broad definition of exposure assessed through questionnaires (often not validated).
- Deficiencies in outcome assessment (broad outcome definitions and use of self-reported outcomes or surrogate outcomes).
- Deficiencies in reporting and analysis (interpretation of effect estimates, confounder control and multiple testing).
- Selective reporting, publication bias and other biases (e.g. conflict of interest).

The observed heterogeneity in the results within each studied outcome was often large. However, heterogeneity is not always a result of biases and may be genuine and consideration of *a priori* defined subgroup analysis and meta-regression should be part of evidence synthesis efforts. Occupational studies, which are of particular importance to pesticide exposure, are also vulnerable to the healthy worker effect, a bias resulting in lower morbidity and mortality rates within the workforce than in the general population. The healthy worker effect tends to decline with increasing duration of employment and length of follow-up.

Studies with sufficient statistical power, detailed definition of pesticide exposure, data for many health outcomes and transparent reporting are rare, apart from the Agricultural Health Study (AHS) and other similarly designed studies. It is important to note that several of these methodological limitations have not been limited to pesticide exposure studies and, most importantly, are not specific in epidemiology and have been observed in other specific fields including in animal studies (Tsilidis et al., 2013).

Given the wide range of pesticides with various definitions found in the EFSA External scientific report, it is difficult to harmonise this information across studies. Although heterogeneity of findings across studies can be as informative as homogeneity, information needs to be harmonised such that replication can be assessed and summary effect sizes be calculated. This does not mean that if there is

genuine heterogeneity the different studies cannot be pooled. Limited conclusions can be made from a single study. Nonetheless, the report highlighted a number of associations between pesticides and health effects that merit further consideration and investigation. Of interest is the fact that a considerable proportion of the published literature focused on pesticides no longer approved for use in the EU and in most developed countries e.g. studies focusing solely on DDT and its metabolites constituted almost 10% of the eligible studies (Ntzani et al., 2013). These may still be appropriate since they may persist as pesticide residues or because they continue to be used in developing countries. Also, the report focused on epidemiological evidence in relation to any health outcome across an approximately 5-year window. Although the report is valuable in describing the field of epidemiological assessment of pesticide–health associations, it is not able to answer specific disease-pesticide questions thoroughly. A more in-depth analysis of specific disease endpoints associated with pesticides exposure is needed, where this information is available, and studies published earlier than the time window covered by the EFSA External scientific report should be also included.

3.2. Limitations in study designs

For ethical reasons, randomised controlled trials are not allowed to test the safety of low dose pesticide exposure in the EU. Therefore, information on potential adverse health consequences in humans has to be extracted using observational studies.

For diseases with long-latency periods, measurement of exposure at one time point may not accurately reflect the long-term exposure which is needed to develop such diseases. This is particularly important for non-persistent pesticides, whose levels in biological samples are not constant but vary quite often. Thus, those studies that claim an association between a single measurement in urine samples and a long latency outcome should be carefully interpreted.

Among the 795 studies reviewed in the Ntzani report, 38% were case–control studies and 32% cross-sectional studies. As a result, evidence on potential adverse health consequences of pesticide exposure is largely based on studies that lack prospective design at least for outcomes that have long latency periods. For the cross-sectional studies, directionality cannot be assessed and observed associations may often reflect reverse causation (is the disease caused by the exposure, or does the disease influence the exposure?). Although reverse causation is a potential problem of cross-sectional studies in many fields of epidemiology, in pesticide epidemiology, it is less of an issue, because in most situations it is unlikely that a disease will cause exposure to pesticides.

Although case–control studies are frequently used for rare outcomes, such as several cancers, their main limitation is that they are prone to recall bias and they have to rely on retrospective assessment of exposure. However, they can still provide useful information, especially for rare outcomes. It is important to examine whether results from case–control and prospective studies converge. This was, for example, the case amongst studies that were conducted to examine associations between intake of *trans*-fatty acids and cardiovascular disease (EFSA, 2004), where both case–control and prospective studies consistently reported positive associations. The effect estimates between the two study designs were systematically different with prospective studies reporting more modest effect sizes but both study designs reached similar conclusions. As for pesticides, similar values have been observed for the magnitude of association between Parkinson's disease and pesticide exposure irrespective of the study design (reviewed in Hernández et al., 2016).

3.3. Relevance of study populations

Because the environmentally relevant doses of pesticides to which individuals are exposed are lower than those required to induce observed toxicity in animal models, the associated toxic effects need to be understood in the context of differences of susceptibility of subpopulations. Potentially vulnerable groups are at an increased risk against exposure to low levels of pesticides than healthy individuals, sometimes during sensitive windows of exposure. This is the case of genetic susceptibility, which represents a critical factor for risk assessment that should be accounted for (Gómez-Martín et al., 2015). Genetic susceptibility largely depends on functional genetic polymorphisms affecting toxicokinetics (e.g. genes encoding xenobiotic metabolising enzymes and membrane transporters) and/or toxicodynamics (e.g. different receptor gene polymorphisms). This genetic variability should be considered on the basis of a plausible scientific hypothesis.

While different disorders, particularly neurodegenerative diseases (Parkinson's disease, Alzheimer's disease, amyotrophic lateral sclerosis) have been linked to exposures to environmental factors (e.g.

pesticides), in many instances the genetic architecture of the disorder has not been taken into account. The prevalence of specific gene mutations may reach 5–10% and sometimes over 20% of cases in certain populations (Gibson et al., 2017), so that the links of these diseases to pesticide exposure may be heavily influenced by genetic structure within populations under study. Given the small effect sizes for many of these disorders, the underlying effects of specific genes not accounted for in the study design may modify the disease risk estimates. Hence, associations with pesticide exposure may need to be evaluated in the light of common genetic influences known to be associated with a spectrum of neurodegenerative diseases. However, genetic variation by itself does not predispose people for an increased pesticide exposure.

A subgroup of population of special interest is represented by children, because their metabolism, physiology, diet and exposure patterns to environmental chemicals differ from those of adults and can make them more susceptible to their harmful effects. The window(s) of biologic susceptibility remain unknown for the most part, and would be expected to vary by mechanism. Gender-based susceptibility also merits consideration in case of pesticide-related reproductive toxicity and endocrine disruption. Those subgroups are currently considered during the risk assessment process but may deserve more attention to provide additional protection.

3.4. Challenges in exposure assessment

The main limitations of epidemiological studies conducted on pesticides derive from uncertainty in exposure assessment. Limitations include the fact that most currently approved pesticides tend to have short elimination half-lives and that their use involves application of various formulations depending on the crop and season. As a result, accurate assessment needs to capture intermittent long-term exposure of these non-persistent chemicals as well as being able to quantify exposure to individual pesticides.

Numerous studies have assessed internal exposure by measuring urinary non-active metabolites common for a large group of pesticides (for example, dialkyl phosphates for organophosphates, 3-phenoxybenzoic acid for pyrethroids or 6-chloronicotinic acid for neonicotinoids). These data should not be utilised to infer any risk because: (a) a fraction of these metabolites might reflect direct exposure through ingestion of preformed metabolites from food and other sources, rather than ingestion of the parent compound and (b) the potency of the different parent pesticides can vary by orders of magnitude. Thereby, HBM data based on those urine metabolites can be unhelpful unless they are paired with other data indicating the actual pesticide exposure.

Ideally exposure should be quantified on an individual level using biomarkers of internal dose. As most available biomarkers reflect short term (few hours or days) exposure and given the cost and difficulty of collecting multiple samples over time, many studies quantify exposure in terms of external dose. Quantitative estimation of external dose needs to account for both frequency and duration of exposure and should preferably be done on an individual but not group level. Often external exposure is quantified using proxy measures such as:

- subject- or relative-reported jobs, job titles, tasks or other lifestyle habits which are being associated with the potential exposure to or actual use of pesticides in general;
- handling of a specific product or set of products and potential exposure to these as documented through existing pesticide records or diaries or estimated from crops grown;
- environmental data: environmental pesticide monitoring, e.g. in water, distance from and/or duration of residence in a particular geographical area considered to be a site of exposure.

In many cases, these proxy measures are recorded with use of questionnaires, which can be either interviewer-administered or based on self-report. However, questionnaire data often rely on individual recall and knowledge and are thus potentially subject to both recall bias and bias introduced by the interviewer or study subjects. These sources of bias can to some extent be quantified if the questionnaires are validated against biomarkers (that is, to what extent do individual questions predict biomarker concentrations in a sub-sample of participants). If the exposure is assessed retrospectively the accuracy of the recall is for obvious reasons more likely to be compromised and impossible to validate. When exposure is based on records, similar difficulties may occur due to, e.g. incomplete or inaccurate records.

In many previous studies, duration of exposure is often used as a surrogate of cumulative exposure, assuming that exposure is uniform and continuous over time (e.g. the employment period) but this assumption must be challenged for pesticides. Although for some chemicals the exposure patterns may be fairly constant, exposures for the large number of pesticides available in the market

will vary with season, by personal protective equipment (PPE) and by work practices, and in many cases, uses are not highly repetitive. At an individual level, exposures can vary on a daily and even hourly basis, and often involve several pesticides. This temporal variability can result in particularly high variation in systemic exposures for pesticides with short biological half-lives and considerable uncertainty in extrapolating single or few measurements to individual exposures over a longer term. Hence, many repeated measurements over time may be required to improve exposure estimates.

3.5. Inappropriate or non-validated surrogates of health outcomes

Self-reported health outcomes are frequently used in epidemiological research because of the difficulty of verifying responses in studies with large samples and limited funds, among other reasons. Although a number of studies have examined agreement between self-reported outcomes and medical records, the lack of verification of such metrics can lead to misclassification, particularly in large population-based studies, which may detract from reliability of the associations found.

Reliance on clinically manifested outcomes can increase the likelihood that individuals who have progressed along the toxicodynamic continuum from exposure to disease but have not yet reached an overt clinical disease state will be misclassified as not having the disease (Nachman et al., 2011). Thereby, delay in onset of clinical symptoms following exposure may cause underreporting where clinical assessment alone is used at an inappropriate point in time.

In the case of carcinogenesis, there are some examples where subclinical outcomes have been assessed as preneoplastic lesions with potential to progress to neoplastic conditions. This is the case of monoclonal gammopathy of undetermined significance (MGUS), which has been associated with pesticide exposure in the AHS (Landgren et al., 2009), as this condition has a 1% average annual risk of progression to malignant multiple myeloma (Zingone and Kuehl, 2011). However, it is difficult to predict if and when an MGUS will progress to multiple myeloma. Since there are studies indicating that pesticide exposure may be associated with the risk of precancerous lesions in animal research, a combined epidemiological analysis of both preneoplastic and neoplastic outcomes may increase the power of such an analysis.

Surrogate outcomes may seem an attractive alternative to clinically relevant outcomes since there may be various surrogates for the same disease and they may occur sooner and/or be easier to assess, thereby shortening the time to diagnosis. A valid surrogate endpoint must, however, be predictive of the causal relationship and accurately predict the outcome of interest. In addition, these surrogates should be relevant to the mode of action of a pesticide such that they should be anchored to established toxicological endpoints to support their predictivity. Although surrogate markers may correlate with an outcome, they may not capture the effect of a factor on the outcome. This may be because the surrogate may not be causally or strongly related to the clinical outcome, but only a concomitant factor, and thus may not be predictive of the clinical outcome. The validity of surrogate outcomes may thus represent a major limitation to their use (la Cour et al., 2010).

However, concerns arise as to whether critical regulatory decisions can be made based on epidemiological studies that did not directly measure the adverse health outcome but valid surrogates instead. The use of surrogates as replacement endpoints should be considered only when there is substantial evidence to establish their reliability in predicting clinical meaningful effects.

3.6. Statistical analyses and interpretation of results

The statistical analyses and the interpretation of scientific findings that appear in the epidemiological literature on the relationship between pesticides and health outcomes do not substantially deviate from those reported in other fields of epidemiological research. Therefore, the advantages and limitations of epidemiological studies presented in Section 2.5 also apply to the epidemiological studies on pesticides.

The few distinctive features of the epidemiological studies on pesticides include the following: (a) sparse use of appropriate statistical analyses in the presence of measurement errors when assessing exposure to pesticides and (b) paucity of information on other important factors that may affect the exposure–health outcome relationship. These features are expanded on in the following paragraphs.

a) Statistical analyses in the presence of measurement errors

The difficulties inherent in correctly measuring exposure are frequent in many areas of epidemiological research, such as nutritional epidemiology and environmental epidemiology. It is not

easy to gauge the short- and long-term exposure outside controlled laboratory experimental settings. In large populations, individuals are exposed to a variety of different agents in a variety of different forms for varying durations and with varying intensities.

Unlike nutritional or environmental epidemiology, however, pesticide epidemiology has so far made little use of statistical analyses that would appropriately incorporate measurement errors, despite their wide availability and sizable literature on the topic. A direct consequence of this is that the inferential conclusions may not have been as accurate and as precise as they could have been if these statistical methods were utilised (Bengtson et al., 2016; Dionisio et al., 2016; Spiegelman, 2016).

b) Information on other important factors of interest

Identifying and measuring the other relevant factors that might affect an outcome of interest is a recurrent and crucial issue in all fields of science. For example, knowing that a drug effectively cures a disease on average may not suffice if such drug is indeed harmful to children or pregnant women. Whether or not age, pregnancy and other characteristics affect the efficacy of a drug is an essential piece of information to doctors, patients, drug manufacturers and drug-approval agencies alike.

Pesticide epidemiology provides an opportunity for careful identification, accurate measuring and thorough assessment of possible relevant factors and their role in the exposure–health outcome relationship. Most often, relevant factors have been screened as potential confounders. When confounding effects were detected, these needed to be adjusted for in the statistical analyses. This has left room for further investigations that would shed light on this important issue by reconsidering data that have already been collected and that may be collected in future studies. The statistical methods in the pesticide literature have been mainly restricted to standard applications of basic regression analyses, such as binary probability and hazard regression models. Potentially useful analytical approaches, such as propensity score matching, mediation analyses, and causal inference, would be helpful for pesticide epidemiology (Imbens and Rubin, 2015).

4. Proposals for refinement to future epidemiological studies for pesticide risk assessment

This section is aimed at addressing methods for assessment of available pesticide epidemiological studies and proposals for improvement of such studies to be useful for regulatory purposes.

When considering the potential regulatory use of epidemiological data, many of the existing epidemiological studies on pesticides exposure and health effects suffer from a range of methodological limitations or deficiencies which limit their value in the assessment of individual active substances. Epidemiological studies on pesticides exposure and health effects would ideally generate semi-quantitative data or be able to have greater relevance to quantitative risk assessment with respect to the output from prediction models. This would allow epidemiological results to be expressed in terms more comparable to the quantitative risk assessments, which are more typically used in evaluating the risks of pesticides. The question arises how such epidemiological data could be considered for risk assessment when judged in comparison to the predictive models. A precisely measured quantitative dose–response relationship is presently rarely attainable as a result of current pesticide epidemiological studies.

The quality, reliability and relevance of the epidemiological evidence in relation to pesticide exposure and health effects can be enhanced by improving (a) the quality of each individual study and (b) the assessment of the combined evidence accrued from all available studies.

4.1. Assessing and reporting the quality of epidemiological studies

The quality and relevance of epidemiological research should be considered when selecting epidemiological studies from the literature for use in risk assessment. The quality of this research can be enhanced by (US-EPA, 2012; Hernández et al., 2016):

a) an adequate assessment of exposure, preferentially biomarker concentrations at individual level reported in a way which will allow for a dose–response assessment;
b) a reasonably valid and reliable outcome assessment (well-defined clinical entities or validated surrogates);
c) an adequate accounting for potentially confounding variables (including exposure to multiple chemicals);
d) the conduct and reporting of subgroup analysis (e.g. stratification by gender, age, ethnicity).

It is widely accepted that biomedical research is subject to and suffers from diverse limitations. An assessment of weaknesses in the design, conduct and analysis of epidemiology research studies on pesticides is essential to identify potentially misleading results and identify reliable data.

Guidelines and checklists help individuals meet certain standards by providing sets of rules or principles that guide towards the best behaviour in a particular area. Several tools and guidelines have been developed to aid the assessment of epidemiological evidence; however, there is no specific tool for assessing studies on pesticides. Although these studies have special considerations around exposure assessment that require specific attention, standard epidemiological instruments for critical appraisal of existing studies may apply. Existing reporting guidelines usually specify a minimum set of information needed for a complete and clear account of what was done and what was found during a research study focusing on aspects that might have introduced bias into the research (Simera et al., 2010).

A number of tools were specifically designed for quality appraisal of observational epidemiological studies, such as the Newcastle–Ottawa scale (NOS) and the Research Triangle Institute (RTI) item bank. The latter is a practical and validated tool which consists of a checklist of 29 questions for evaluating the risk of bias and precision of epidemiological studies of chemical exposures. In addition, the Biomonitoring, Environmental Epidemiology, and Short-Lived Chemicals (BEES-C) instrument was developed to evaluate the quality of epidemiological research that use biomonitoring to assess short-lived chemicals (LaKind et al., 2015), but it can also be used for persistent chemicals and environmental measures as its main elements are cross-cutting and are more broadly applicable. Two earlier efforts to develop evaluative schemes focused on epidemiology research on environmental chemical exposures and neurodevelopment (Amler et al., 2006; Youngstrom et al., 2011).

Regarding quality of reporting, the Enhancing the QUAlity and Transparency Of health Research (EQUATOR) Network, officially launched in June 2008, is an international initiative that promotes transparent and accurate reporting of health research studies. It currently lists over 90 reporting guidelines with some of them being specific for observational epidemiological studies (e.g. Strengthening the Reporting of OBservational studies in Epidemiology (STROBE)). The STROBE statement includes recommendations on what should be included in an accurate and complete report of an observational study including cross-sectional, case–control and cohort studies using a checklist of 22 items that relate to the title, abstract, introduction, methods, results and discussion sections of articles (von Elm et al., 2007). The STROBE statement has been endorsed by a growing number of biomedical journals which refer to it in their instructions for authors. Table 1 presents a summary of the main features that STROBE proposes to be taking into account when assessing the quality of reporting epidemiological studies. Extensions to STROBE are available including the STROBE Extension to Genetic Association studies (STREGA) initiative and the STROBE-ME statement for assessment of molecular epidemiology studies. Since the STROBE checklist mentions only in a general way exposure and health outcomes, the PPR Panel recommends that an extension of the STROBE statement be developed, for inclusion in the EQUATOR network library, specifically relevant to the area of pesticide exposure and health outcomes. This would greatly assist researchers and regulatory bodies in the critical evaluation of study quality.

Table 1: Main features of the STROBE tool to assess quality of reporting of epidemiological studies

STROBE Statement Items		
Factor	**Item**	**Recommendation**
Title and Abstract		
	1	a) Indicate the study's design with a commonly used term in the title of the abstract b) Provide in the abstract an informative and balanced summary of what was done and what was found
Introduction		
Background/rationale	2	Explain the scientific background and rationale for the investigation being reported
Objectives	3	State specific objectives, including prespecified hypotheses
Methods		
Study design	4	Present key elements of study design early in the paper
Setting	5	Describe the setting, locations and relevant dates, including periods of recruitment, exposure, follow-up and data collection

STROBE Statement Items		
Factor	**Item**	**Recommendation**
Participants	6	a) Cohort study – Give the eligibility criteria, and the sources and methods of selection of participants. Describe methods of follow-up Case–control study – Give the eligibility criteria, and the sources and methods of case ascertainment and control selection. Give the rationale for the choice of cases and controls Cross-sectional study – Give eligibility criteria, and the sources and methods of selection of participants b) Cohort study – For matched studies, give matching criteria and the number of exposed and unexposed Case–control study – For matched studies, giving matching criteria and the number of controls per case
Variables	7	Clearly define all outcomes, exposures, predictors, potential confounders and effect modifiers. Give diagnostic criteria, if applicable
Data sources/ measurements	8*	For each variable of interest, give sources of data and details of methods of assessment (measurement). Describe comparability of assessment methods if there is more than one group
Bias	9	Describe any efforts to address potential sources of bias
Study size	10	Explain how the study size was arrived at
Quantitative variables	11	Explain how quantitative variables were handled in the analyses. If applicable, describe which groupings were chosen and why
Statistical methods	12	a) Describe all statistical methods, including those used to control for confounding b) Describe any methods used to examine subgroups and interactions c) Explain how missing data were addressed d) Cohort study – If applicable, explain how loss to follow-up was addressed Case–control study – If applicable, explain how matching of cases and controls was addressed Cross-sectional study – If applicable, describe analytical methods taking account of sampling strategy e) Describe any sensitivity analyses
Results		
Participants	13*	a) Report numbers of individuals at each stage of study – e.g. numbers potentially eligible, examined for eligibility, confirmed eligible, included in the study, completing follow-up and analysed b) Give reasons for non-participation at each stage c) Consider use of a flow diagram
Descriptive data	14*	a) Give characteristics of study participants (e.g. demographic, clinical, social) and information on exposures and potential confounders b) Indicate number of participants with missing data for each variable of interest c) *Cohort study* – Summarise follow-up time (e.g. average and total amount)
Outcome data	15*	*Cohort study* – Report numbers of outcome events or summary measures over time *Case–control study* – Report numbers in each exposure category, or summary measures of exposure *Cross-sectional study* – Report numbers of outcome events or summary measures
Main results	16	a) Give unadjusted estimates and, if applicable, confounder-adjusted estimates and their precision (e.g. 95% confidence interval). Make clear which confounders were adjusted for and why they were included b) Report category boundaries when continuous variables were categorised c) If relevant, consider translating estimates of relative risk into absolute risk for a meaningful time period

STROBE Statement Items		
Factor	**Item**	**Recommendation**
Other analyses	17	Report other analyses done – e.g. analyses of subgroups and interactions, and sensitivity analyses
Discussion		
Key results	18	Summarise key results with reference to study objectives
Limitations	19	Discuss limitations of the study, taking into account sources of potential bias or imprecision. Discuss both direction and magnitude of any potential bias
Interpretation	20	Give a cautious overall interpretation of results considering objectives, limitations, multiplicity of analyses, results from similar studies, and other relevant evidence
Generalisability	21	Discuss the generalisability (external validity) of the study results
Other information		
Funding	22	Give the source of funding and the role of the funders for the present study and, if applicable, for the original study on which the present article is based

*: Give information separately for cases and controls in case–control studies and, if applicable, for exposed and unexposed groups in cohort and cross-sectional studies.

Selective reporting can occur because non-significant results or unappealing significant results may not be published. Investigators should avoid the selective reporting of significant results and high-risk estimates. In this regard, standardisation of reporting of epidemiological studies could help to reduce or avoid selective reporting. The STROBE statement and similar efforts are useful tools for this purpose. Although some epidemiological research will remain exploratory and *post hoc* in nature, this should be clarified in the publications and selective reporting minimised, so that epidemiological findings could be interpreted in the most appropriate perspective (Kavvoura et al., 2007).

Preregistration of studies and prepublication of protocols are the measures taken by some Journal editors and Ethics Committees to reduce reporting bias and publication bias in clinical trials on pharmaceuticals. Although a similar proposal has been suggested for observational epidemiological studies in order to be conducted as transparently as possible to reduce reporting bias and publication bias, there is no consensus among epidemiologists (Pearce, 2011; Rushton, 2011). In contrast, a number of initiatives have been undertaken by professional societies to foster good epidemiological practice. This is the case, for example, of the International Epidemiological Association (IEA, 2007) or the Dutch Society for Epidemiology on responsible epidemiologic Research Practice (DSE, 2017).

Data quality assessment of formal epidemiological studies is based solely on the methodological features of each individual study rather than on the results, regardless of whether they provide evidence for or against an exposure/outcome association. However, for risk assessment, it is important to assess not only the quality of study methods but also the quality of the information they provide. Indeed, good studies may be dismissed during the formal quality assessment by the poor reporting of the information.

4.2. Study design

Well conducted prospective studies with appropriate exposure assessment provide the most reliable information and are less prone to biases. When prospective studies are available, results from studies of less robust design can give additional support. In the absence of prospective studies the results from cross-sectional and case–control studies should be considered but interpreted with caution. However, it is acknowledged that a well-designed case–control study may be superior to a less well designed cohort study. Analytical approaches should be congruent with the study design, and assumptions that the statistical methods required should be carefully evaluated.

Ideally observational studies for long-term diseases should be prospective and designed such that the temporal separation between the exposure and the health outcome is appropriate with respect to the time it takes to develop the disease. For outcomes such as cancer or cardiovascular diseases, which often have a long latency period (> 10 years), exposure should be assessed more than once prior to the outcome assessment. For other outcomes with a shorter latency period, such as immune function disturbances, the appropriate temporal separation may be in the range of days or weeks and a single exposure assessment may be adequate. In short, the ideal design of a study depends on the latency period for the outcome under consideration. The expected latency period then determines both the length of follow-up and the frequency for which the exposure has to be quantified.

4.3. Study populations

The EU population, which exceeds 500 million people, can be assumed to be fairly heterogeneous and so expected to include a number of more sensitive individuals that may be affected at lower doses of pesticide exposure. To address this, in stratified sampling, the target population is divided into subgroups following some key population characteristics (e.g. sex, age, geographic distribution, ethnicity or genetic variation) and a random sample is taken within each subgroup. This allows subpopulations to be represented in a balanced manner in the study population.

Vulnerable populations should then be examined in epidemiological studies either through subgroup or sensitivity analysis. However, such analyses need to be defined *a priori*. In case of ad hoc subgroup sensitivity analysis, the statistical thresholds should be adjusted accordingly and the replication of results should follow. Evidence of vulnerable subpopulations would ideally involve prospective studies that include assessment of biomarkers of exposure, subclinical endpoints and disease incidence over time.

It may be impossible to find a threshold of a toxic-induced increase in disease in the population because a large number of people are in a preclinical state and would be sensitive to the low end of the dose–response curve. For that to be evident, the epidemiology data would need to characterise the relationship between chemical exposure and risk of disease in a broad cross-section of the population (or look at precursor lesions or key events) and allow a robust examination of a low-dose slope.

On the basis of the degree of evidence relevant to a vulnerable subpopulation, consideration should be given to whether dose–response assessment will focus on the population as a whole or will involve separate assessments for the general population and susceptible subgroups. If it is the population as a whole, the traditional approach is to address variability with uncertainty factors; it may also be possible to analyse the effect of variability on risk by evaluating how the risk distribution of the disease shifts in response to the toxicant. In essence, the risk distribution based on a subclinical biomarker is an expression of toxicodynamic variability that can be captured in dose–response assessment.

The alternative approach is to address vulnerable subpopulations as separate from the general population and assign them unique potencies via dose–response modelling specific to the groups that might be based on actual dose–response data for the groups, on adjustments for specific toxicokinetic or toxicodynamic factors, or on more generic adjustment or uncertainty factors. For a pesticide, if it is known that a particular age group, disease (or disease-related end-point), genetic variant or co-exposure creates unique vulnerability, efforts should be made to estimate the potency differences relative to the general population and on that basis to consider developing separate potency values or basing a single value on the most sensitive group or on the overall population with adjustments for vulnerable groups.

4.4. Improvement of exposure assessment

The difficulties often associated with pesticide exposure assessment in epidemiological studies have been highlighted above. The description of pesticide exposure (in particular quantitative information on exposure to individual pesticides) is generally reported in insufficient detail for regulatory purposes and this limitation is difficult to overcome, especially for diseases with a long latency period (e.g. many cancers and neurodegenerative disorders).

It is noteworthy that the methods necessary to conduct exposure monitoring are to be submitted by the applicant in the dossier. The regulation requirements do ask for validated methods that can be used for determining exposure. The Commission Regulation (EU) No 283/2013, setting out the data requirements for active substances, in accordance with Regulation (EC) No 1107/2009 of the European Parliament and of the Council concerning the placing of PPP on the market, addresses information on methods of analysis required to support both pre-approval studies and post-approval monitoring. In this context, the post-approval requirements are the most relevant and the regulation literally states:

'4.2. Methods for post-approval control and monitoring purposes – Methods, with a full description, shall be submitted for:

a) the determination of all components included in the monitoring residue definition as submitted in accordance with the provisions of point 6.7.1 in order to enable Member States to determine compliance with established maximum residue levels (MRLs); they shall cover residues in or on food and feed of plant and animal origin;

b) the determination of all components included for monitoring purposes in the residue definitions for soil and water as submitted in accordance with the provisions of point 7.4.2;
c) the analysis in air of the active substance and relevant breakdown products formed during or after application, unless the applicant shows that exposure of operators, workers, residents or bystanders is negligible;
d) the analysis in body fluids and tissues for active substances and relevant metabolites.

As far as practicable these methods shall employ the simplest approach, involve the minimum cost, and require commonly available equipment. The specificity of the methods shall be determined and reported. It shall enable all components included in the monitoring residue definition to be determined. Validated confirmatory methods shall be submitted if appropriate. The linearity, recovery and precision (repeatability) of methods shall be determined and reported.

Data shall be generated at the LOQ and either the likely residue levels or ten times the LOQ. The LOQ shall be determined and reported for each component included in the monitoring residue definition. For residues in or on food and feed of plant and animal origin and residues in drinking water, the reproducibility of the method shall be determined by means of an independent laboratory validation (ILV) and reported'.

From this, it can be concluded that the requirements exist, but are somewhat less stringent for human biomonitoring than for monitoring of residues in food and feed.

Failure to use these existing methods restricts the potential for the use of epidemiological evidence in the regulation of specific pesticides. It is therefore important that those contemplating future studies carefully consider approaches to be used to avoid misclassification of exposure, and to conduct appropriate detailed exposure assessments for specific pesticides, which allow for sound dose–response analyses, and demonstrate the validity of the methods used.

A given exposure may have a different health impact depending on the period in the lifespan when exposure takes place. Greater attention needs to be paid to exposures occurring during periods of potential susceptibility for disease development by ensuring that the exposure assessment adequately addresses such critical times. This may be particularly relevant for studies involving neurodevelopment, obesity or allergic responses, which are complex multistage developmental processes that occur either prenatally or in the early post-natal life. For this reason, measurement of the exposure at one single time period may not properly characterise relevant exposures for all health effects of the environmental factors, and thus, the possibility arises of needing to measure the exposure at several critical periods of biological vulnerability to environmental factors. It is particularly challenging to construct an assessment of historical exposures which may deviate from current exposures, in both the range of chemicals and intensity of exposure and also co-exposure to other substances which are not included in the scope of study.

There are advantages and disadvantages to all methods of measuring pesticide exposure, and specific study designs and aims should be carefully considered to inform a specific optimal approach.

Exposure assessment can be improved at the *individual* level in observational research by using:

a) **Personal exposure monitoring:** This can be used to document exposures as readings measure pesticide concentration at the point of contact. Personal exposure monitors have been costly and burdensome for study participants. However, technological advances have recently driven personal exposure monitoring for airborne exposures to inexpensive, easy to use devices and these are suitable for population research. Personal exposure monitors that are specific to pesticide exposure could involve sensors to measure airborne concentrations, 'skin' patches to measure dermal concentrations, indoor home monitors that capture dust to measure other means of exposure. These mobile technology advances can be employed to provide observational studies with detailed and robust exposure assessments. Such equipment is now increasingly being adapted to serve large-scale population research and to capture data from large cohort studies. These coupled with other technological advances, such as real time data transfers via mobile phones and mobile phone applications to capture lifestyle and other habits, could bring next generation observational studies far more detailed and robust exposure assessments compared to current evidence. However, the generation of huge volumes of data can pose organisational, statistical and technical challenges, particularly with extended follow-up times. Ethics and personal data protection issue should be taken into account, and local regulations may prevent extensive use of such technologies. However, use of such personal monitors only provides information for one of the different potential routes of exposure.

b) **Biomarkers of exposure** (human biomonitoring (HBM)). An alternative and/or complementary approach is to ascertain the internal dose, which is the result of exposure via different routes (dermal,

inhalation and dietary exposure). These biomarkers have the potential to play an important role in assessing aggregate exposure to pesticides and informing cumulative risk assessment. Biomonitoring requires measurements in biological samples of concentrations of chemical under consideration (parent or metabolites) or markers of pathophysiologic effects thereof (such as adducts). However, challenges may include uncertainties relating to extrapolation of measured concentrations in biological samples to relevant doses.

Although biomonitoring has the potential to provide robust estimates of absorbed doses of xenobiotics, modern pesticides and their metabolites are eliminated from the body relatively quickly, with excretion half-lives typically measured in a few days (Oulhote and Bouchard, 2013). Consequently, use of biomarkers is both resource intensive and intrusive. The process is even more intrusive when it has to be conducted repeatedly on large numbers of individuals to monitor exposures over long durations.

Nevertheless, because of the potential to provide accurate integrated estimates of absorbed doses, biological monitoring of pesticides and their metabolites can be usefully employed to calibrate other approaches of exposure assessment. A good example of such an approach is that used by the Agricultural Health Study (Thomas et al., 2010; Coble et al., 2011; Hines et al., 2011). Also, HBM methods can be used with other forms of exposure assessment for the construction of long exposure histories.

Biomonitoring improves the precision in characterisation of exposure and allows the investigation of changes in exposure that occur at environmentally relevant exposure concentrations. Data collected in large-scale biomonitoring studies can be useful in setting reference ranges to assist in exposure classification in further epidemiological studies. Biomonitoring data also provide critical information for conducting improved risk assessment and help to identify subpopulations at special risk for adverse outcomes.

Biobanks, as repositories of biological samples, can be exploited to assess biomarkers of exposure with the aim of investigating early exposure–late effect relationships. That is, whether exposures occurring during early life are critical for disease development later in life (e.g. neurobehavioral impairment, children tumours, immunotoxic disorders, etc.) and to retrospectively assess health risks according to current health guidelines.

The results of measurements of metabolite levels in human matrices, e.g. urine, blood or hair do not provide the complete story with respect to the actual received dose. Additional assessment, possibly employing physiological-based toxicokinetic (PBTK) approaches, may be required to estimate the total systemic or tissue/organ doses. A PBTK model is a physiologically based compartmental model used to characterise toxicokinetic behaviour of a chemical, in particular for predicting the fate of chemicals in humans. Data on blood flow rates, metabolic and other processes that the chemical undergoes within each compartment are used to construct a mass-balance framework for the PBTK model. PBTK models cannot be used only to translate external exposures into an internal (target) dose in the body, but also to infer external exposures from biomonitoring data. Furthermore, PBTK models need to be validated.

Toxicokinetic processes (ADME) determine the 'internal concentration' of an active substance reaching the target and help to relate this concentration/dose to the observed toxicity effect. Studies have been prescribed by the current regulations, but it would be beneficial to survey all the evidence, be it from *in vitro*, animal or human studies, about toxicokinetic behaviour of an active substance. Further discussion on quality assurance issues and factors to consider in relation to HBM studies is present in the report of the EFSA outsourced project (Bevan et al., 2017).

Exposure assessment can also be improved at the *population* level in observational research by using:

a) Larger epidemiological studies that make use of novel technologies and big data availability, such as **registry data** or data derived from large databases (including administrative databases) on health effects and pesticide usage, could provide more robust findings that might eventually be used for informed decision-making and regulation. Much effort needs to concentrate around the use of registered data which may contain records of pesticide use by different populations, such as farmers or other professional users that are required to maintain.[9] Such data could be further linked to

[9] Regulation 1107/2009 Article 67 states: Record-keeping 1. Producers, suppliers, distributors, importers, and exporters of plant protection products shall keep records of the plant protection products they produce, import, export, store or place on the market for at least 5 years. Professional users of plant protection products shall, for at least 3 years, keep records of the plant protection products they use, containing the name of the plant protection product, the time and the dose of application, the area and the crop where the plant protection product was used. They shall make the relevant information contained in these records available to the competent authority on request. Third parties such as the drinking water industry, retailers or residents, may request access to this information by addressing the competent authority. The competent authorities shall provide access to such information in accordance with applicable national or Community law.

electronic health records (*vide supra*) and provide studies with unprecedented sample size and information on exposure and subsequent disease and will eventually be able to answer robustly previously unanswered questions. At the same time, information on active substances needs to be better captured in these registries and large databases. Dietary pesticide residue exposure can be estimated more accurately by using spraying journal data in combination with supervised residue trials. This method has the advantage of including more comprehensive and robust source data, more complete coverage of used pesticides and more reliable and precise estimates of residues below standard limit of quantification (LOQ) (Larsson et al., 2017).

b) Novel sophisticated approaches to **geographical information systems** (GIS) and small area studies might also serve as an additional way to provide estimates of residential exposures. Exposure indices based on GIS (i.e. residential proximity to agricultural fields and crop surface with influence around houses), when validated, may represent a useful complementary tool to biomonitoring and have been used to assess exposure to pesticides with short biological half-lives (Cornelis et al., 2009). As some such exposures maybe influenced by wind direction, amongst other factors, this should be taken into account through a special analysis of outcomes to make best of use of the approach. Also, these indices could be more representative, albeit non-specific, measures of cumulative exposure to non-persistent pesticides for long periods of time than biomonitoring data (González-Alzaga et al., 2015).

As already discussed, to be useful for the regulatory risk assessments of individual compounds epidemiological exposure assessments should provide information on specific pesticides. However, epidemiological studies which include more generic exposure assessments also have the potential to identify general risk factors and suggest inferences of causal associations in relevant human populations. Such observations may be important both informing overall regulatory policies, and for identification of matters for further epidemiological research.

Recent advances in modern technologies make it possible to estimate pesticide exposures to an unprecedented extent using novel analytical strategies:

a) The development of the so called -**omic techniques**, such as metabolomics and adductomics, also presents intriguing possibilities for improving exposure assessment through measurement of a wide range of molecules, from xenobiotics and metabolites recorded over time in biological matrices (blood, saliva, urine, hair, nails, etc.), to covalent complexes with DNA and proteins (adductomics) and understanding biological pathways. These methodologies could be used in conjunction with other tools. There is also both interest and the recognition that further work is required before such techniques can be applied in regulatory toxicology. The use of the exposome (the totality of exposures received by an individual during life) might be better defined by using 'omics' technologies and biomarkers appropriate for human biomonitoring. Nevertheless, important limitations have to be acknowledged because of the lack of validation of these methodologies and their cost, which limits their use at large scale.

b) Environmental exposures are traditionally assessed following 'one-exposure-one-health-effect' approach. In contrast, the **exposome** encompass the totality of human environmental exposures from conception onward complementing the genetics knowledge to characterise better the environmental components in disease aetiology. As such, the exposome includes not only any lifetime chemical exposures but also other external and or internal environmental factors, such as infections, physical activity, diet, stress and internal biological factors (metabolic factors, gut microflora, inflammation and oxidative stress). A complete exposome would have to integrate many external and internal exposures from different sources continuously over the life course. However, a truly complete exposome will likely never be measured. Although all these domains of the exposome need to be captured by using different approaches than the traditional ones, it is envisaged that no single tool will be enough to this end.

The more holistic approach of exposure is not intended to replace the traditional 'one-exposure-one-health-effect' approach of current epidemiological studies. However, it would improve our understanding of the predictors, risk factors and protective factors of complex, multifactorial chronic diseases. The exposome offers a framework that describes and integrates, holistically, the environmental influences or exposures over a lifetime (Nieuwenhuijsen, 2015).

Collaborative research and integration of epidemiological or exploratory studies forming large consortia are needed to validate these potential biomarkers and eventually lead to improved exposure assessment. The incorporation of the exposome paradigm into traditional biomonitoring approaches offers a means to improve exposure assessment. Exposome-wide association studies (EWAS) allow to measurement of thousands of chemicals in blood from healthy and diseased people, test for disease

associations and identify useful biomarkers of exposure that can be targeted in subsequent investigations to locate exposure sources, establish mechanisms of action and confirm causality (Rappaport, 2012). After identifying these key chemicals and verifying their disease associations in independent samples of cases and controls, the chemicals can be used as biomarkers of exposures or disease progression in targeted analyses of blood from large populations.

In relation to the exposome concept, the -omics technologies have the potential to measure profiles or signatures of the biological response to the cumulative exposure to complex chemical mixtures. An important advance would be to identify a unique biological matrix where the exposome could be characterised without assessing each individual exposure separately in a given biological sample. The untargeted nature of omics data will capture biological responses to exposure in a more holistic way and will provide mechanistic information supporting exposure-related health effects. Importantly, omics tools could shed light on how diverse exposures act on common pathways to cause the same health outcomes.

While improved exposure assessment increases the power to detect associations, in any individual study it is necessary to maximise the overall power of the study by optimising the balance between the resource used for conducting an exposure assessment for each subject and the total number of subjects.

4.5. Health outcomes

For pesticides, the health outcomes are broad as these chemicals have not shown a particular effect in relation to just one single disease area. For each health outcome, multiple definitions may exist in the literature with a varying degree of validation and unknown reproducibility across different databases, which are limited by the lack of generalisability. A proper definition of a health outcome is critical to the validity and reproducibility of observational epidemiological studies, and the consistency and clarity of these definitions need to be considered across studies. While prospective observational studies have explicit outcome definitions, inclusion and exclusion criteria and standardised data collection, retrospective studies usually rely on identification of health outcomes based largely on coded data, and classification and coding of diseases may change over time. Detailed description of the actual codes used to define key health outcomes and the results of any validation efforts are valuable to future research efforts (Stang et al., 2012; Reich et al., 2013). An example of coded diseases is the ICD-10, which for instance can be used as a tool to standardise the broad spectrum of malignant diseases.

In some surveillance studies, it is preferable to use broader definitions with a higher sensitivity to identify all potential cases and then apply a narrower and more precise definition with a high positive predictive value to reduce the number of false positives and resulting in more accurate cases. In contrast, in formal epidemiological studies, a specific event definition is used and validated to determine its precision; however, the 'validation' does not test alternative definitions, so it is not possible to determine sensitivity or specificity.

Surrogate endpoints should be avoided unless they have been validated. Some criteria to assess the validity of a surrogate outcome include:

- The surrogate has been shown to be in the causal pathway of the disease. This can be supported by the following evidence: correlation of biomarker response to pathology and improved performance relative to other biomarkers; biological understanding and relevance to toxicity (mechanism of response); consistent response across mechanistically different compounds and similar response across sex, strain and species; the presence of dose–response and temporal relationship to the magnitude of response; specificity of response to toxicity; that is, the biomarker should not reflect the response to toxicities in other tissues, or to physiological effects without toxicity in the target organ.
- At least one well conducted trial using both the surrogate and true outcome (Grimes and Schulz, 2005; la Cour et al., 2010). Several statistical methods are used to assess these criteria and if they are fulfilled the validity of the surrogate is increased. However, many times some uncertainty remains, making it difficult to apply surrogates in epidemiological studies (la Cour et al., 2010).

The data on health outcomes over the whole EU is potentially very extensive. If it can be managed effectively, it will open the prospect of greater statistical power for epidemiological studies assessing deleterious effects using very large sample sizes. Necessary prerequisites for these studies which may

detect new subtle effects, chronic effects or effects on subpopulations when stratified are beyond the remit of risk assessment. They include trans-national approaches to health informatics where harmonised diagnostics, data storage and informatics coupled with legally approved access to anonymised personal data for societal benefit are established. Health records should include adequate toxidrome classification. The latter may in turn require improvements in medical and paramedical training to ensure the quality of the input data.

Another opportunity for biological monitoring to be employed is where the investigation involves the so-called biomarkers of effect. That is a quantifiable biochemical, physiological, or other change that, depending on the magnitude, is associated with an established or possible health impairment or disease. Biomarkers of effect should reflect early biochemical modifications that precede functional or structural damage. Thus, knowledge of the mechanism ultimately leading to toxicity is necessary to develop specific and useful biomarkers, and vice versa, an effect biomarker may help to explain a mechanistic pathway of the development of a disease. Such biomarkers should identify early and reversible events in biological systems that may be predictive of later responses, so that they are considered to be preclinical in nature. Advances in experimental -omics technologies will show promise and provide sound information for risk assessment strategies, i.e. on mode of action, response biomarkers, estimation of internal dose and dose–response relationships (DeBord et al., 2015). These technologies must be validated to assess their relevance and reliability. Once validated, they can be made available for regulatory purposes.

5. Contribution of vigilance data to pesticides risk assessment

In addition to the formal epidemiological studies discussed in Sections 2–4, other human health data can be generated from ad hoc reports or as a planned process, i.e. through monitoring systems that have been implemented at the national level by public health authorities or authorisation holders. Consistent with Sections 2–4, this section first reviews how such a monitoring system should operate, what the current situation is regarding the monitoring of pesticides and what recommendations for improvement can be made.

5.1. General framework of case incident studies

A continuous process of collection, reporting and evaluation of adverse incidents has the potential to improve the protection of health and safety of users and others by reducing the likelihood of the occurrence of the same adverse incident in different places at later times, and also to alleviate consequences of such incidents. This obviously also requires timely dissemination of the information collected on such incidents. Such a process is referred to as vigilance.[10]

For example in the EU, the safety monitoring of medicines is known as pharmacovigilance; the pharmacovigilance system operates between the regulatory authorities in Member States, the European Commission and the European Medicines Agency (EMA). In some Member States, regional centres are in place under the coordination of the national Competent Authorities. Manufacturers and health care professionals report incidents to the Competent Authority at the national level, which ensures that any information regarding adverse reactions is recorded and evaluated centrally and also notifies other authorities for subsequent actions. The records are then centralised by the EMA which supports the coordination of the European pharmacovigilance system and provides advice on the safe and effective use of medicines.

5.2. Key limitations of current framework of case incident reporting

Several EU regulations require the notification and/or collection and/or reporting of adverse events caused by pesticides in humans (occurring after acute or chronic exposure in the occupational setting, accidental or deliberate poisoning, etc.). These include:

- Article 56 of EC Regulation 1107/2009 requires that 'The holder of an authorisation for a plant protection product shall immediately notify the Member States [...] In particular, potentially

[10] The concept of survey refers to a single effort to measure and record something, and surveillance refers to repeated standardized surveys to detect trends in populations in order to demonstrate the absence of disease or to identify its presence or distribution to allow for timely dissemination of information. Monitoring implies the intermittent analysis of routine measurements and observations to detect changes in the environment or health status of a population, but without eliciting a response. Vigilance is distinct from surveillance and mere monitoring as it implies a process of paying close and continuous attention, and in this context addresses specifically post marketing events related to the use of a chemical.

harmful effects of that plant protection product, or of residues of an active substance, its metabolites, a safener, synergist or co-formulant contained in it on human health [. . .] shall be notified. To this end the authorisation holder shall record and report all suspected adverse reactions in humans, in animals and the environment related to the use of the plant protection product. The obligation to notify shall include relevant information on decisions or assessments by international organisations or by public bodies which authorise plant protection products or active substances in third countries'.

- Article 7 of EC Directive 128/2009 establishing a framework for Community action to achieve the sustainable use of pesticides requires that: '2. Member States shall put in place systems for gathering information on pesticide acute poisoning incidents, as well as chronic poisoning developments where available, among groups that may be exposed regularly to pesticides such as operators, agricultural workers or persons living close to pesticide application areas. 3. To enhance the comparability of information, the Commission, in cooperation with the Member States, shall develop by 14 December 2012 a strategic guidance document on monitoring and surveying of impacts of pesticide use on human health and the environment'. However, at the time of publishing this scientific opinion, this document has still not been released.

There are three additional regulations that apply, although indirectly, to pesticides and reporting:

- EC Regulation 1185/2009 concerning statistics on pesticides requires that Member States shall collect data on pesticide sales and uses according to a harmonised format. The statistics on the placing on the market shall be transmitted yearly to the Commission and the statistics on agricultural use shall be transmitted every 5 year.
- Article 50 of Regulation (EC) 178/2002, laying down the general principles and requirements of food law, set up an improved and broadened rapid alert system covering food and feed (RASFF). The system is managed by the Commission and includes as members of the network Member States, the Commission and the Authority. It reports on non-authorised occurrences of pesticides residues and food poisoning cases.
- Article 45 (4) of EC Regulation 1272/2008 (CLP Regulation): importers and downstream users placing hazardous chemical mixtures on the market of an EU Member State will have to submit a notification to the Appointed Body/Poison Centre of that Member State. The notification needs to contain certain information on the chemical mixture, such as the chemical composition and toxicological information, as well as the product category to which the mixture belongs. The inclusion of information on the product category in a notification allows Appointed Bodies/Poison Centres to carry out comparable statistical analysis (e.g. to define risk management measures), to fulfil reporting obligations and to exchange information among MS. The product category is therefore not used for the actual emergency health response as such, but allows the identification of exposure or poisoning trends and of possible measures to prevent future poisoning cases. When formally adopted, the new Regulation will apply as of 1 January 2020.

While there are substantial legislative provisions, to this date a single unified EU 'phytopharmacovigilance'[11] system akin to the pharmacovigilance system does not exist for PPP. Rather, a number of alerting systems have been developed within the EU to alert, notify, report and share information on chemical hazards that may pose a risk to public health in Member States. These systems cover different sectors including medicines, food stuffs, consumer products, industrial accidents, notifications under International Health Regulations (IHR) and events detected by EU Poisons Centres and Public Health Authorities. Each of these systems notify and distribute timely warnings to competent authorities, public organisations, governments, regulatory authorities and public health officials to enable them to take effective action to minimise and manage the risk to public health (Orford et al., 2014).

In the EU, information on acute pesticide exposure/incident originates mainly from data collected and reported by Poison Control Centres (PCC's). PCC's collect both cases of acute and chronic exposure/poisoning they are aware of, in the general population and in occupational settings. Cases are usually well-documented and information includes circumstances of exposure/incident, description of the suspected causal agent, level and duration of exposure, the clinical course and treatment and an assessment of the causal relationship. In severe cases, the toxin and/or the metabolites are usually

[11] 'phytovigilance' would refer to a vigilance system for plants; as pesticides are intended to be 'medicines' for crops, the term 'phytopharmacovigilance' is considered to be the more appropriate one here. Furthermore, it is a broad term used in France covering soil, water, air, environment, animal data, etc.

measured in blood or urine. However, follow-up of cases reported to the centres merits further attention to identify potential long-term protracted effects.

There are two key obstacles to using Poison Centres data: official reports from national Poisons Centres are not always publicly available and when they are, there is a large heterogeneity in the format of data collections and coding, and assessment of the causal relationship. Indeed, each Member State has developed its own tools for collection activities resulting in difficulties for comparing and exchanging exposure data. In 2012, the European Commission funded a collaborative research and development project to support the European response to emerging chemical events: the Alerting and Reporting System for Chemical Health Threats, Phase III (ASHTIII) project. Among the various tools and methodologies that were considered, methods to exchange and compare exposure data from European PCC's were developed. As a feasibility study, work-package 5 included the development of a harmonised and robust coding system to enable Member States to compare pesticide exposure data. However, results of a consultation with the PCC community showed that further coordination of data coding and collection activities is supported. It was concluded that more support and coordination is required at the EU and Member States level so that exposures data can be compared between Member States (Orford et al., 2015).

In addition to data collected by PCC's, several Member States have set up programmes dedicated to occupational health surveillance.[12] The purpose of these programmes is to identify the kinds of jobs, types of circumstances and pesticides that cause health problems in workers in order to learn more about occupational pesticide illnesses and injuries and how to prevent them. They are based on voluntary event notification by physicians (sometimes self-reporting by users) of any case of suspected work-related pesticide injury or illness or poisoning. In addition to medical data, information gathered includes data regarding type of crop, mode of application, temperature, wind speed, wearing of personal protection equipment, etc. Once collected, these data are examined and a report is released periodically; they provide a useful support to evaluate the safety of the products under re-registration. These data also highlight emerging problems and allow definition of evidence-based preventive measures for policy-makers. At EU level, the European Agency for Safety and Health at Work (EU-OSHA)[13] has very little in the way of monitoring of occupational pesticide-related illnesses data. In the USA, a programme specifically dedicated to pesticides funded and administered by the National Institute for Occupational Safety and Health (NIOSH) is in operation in a number of States.[14]

In summary, currently human data may be collected in the form of case reports or case series, poison centres information, coroner's court findings, occupational health surveillance programmes or post-marketing surveillance programmes. However, not all this information is present in the medical data submitted by applicants mainly because the different sources of information are diverse and heterogeneous by nature, which makes some of them sometimes not accessible.

- Data collected through occupational health surveillance of the plant production workers or if they do so, the medical data are quite limited being typically basic clinical blood measurements, physical examinations, potentially with simple indications of how and where exposed took place, and there usually is no long-term follow up. Furthermore, worker exposures in modern plants (especially in the EU) are commonly very low, and often their potential exposure is to a variety of pesticides (unless it is a facility dedicated to a specific chemical).
- Moreover, the reporting of data from occupational exposure to the active substances during manufacture is often combined with results from observations arising from contact with the formulated plant protection product as the latter information results from case reports on poisoning incidents and epidemiological studies of those exposed as a result of PPP use. Indeed, the presence of co-formulants in a plant protection product can modify the acute toxicological profile. Thus, to facilitate proper assessment, when reporting findings collected in humans it should be clearly specified whether it refers to the active substance per se or a PPP.

With regard to the requirements of specific data on diagnoses of poisoning by the active substance or formulated plant protection products and proposed treatments, which are also part of chapter 5.9 of the EC Regulation 283/2013, information is often missing or limited to those cases where the toxic mode of action is known to occur in humans and a specific antidote has been identified.

[12] For example: Phyt'attitude in France is a vigilance programme developed by the Mutualité Sociale Agricole: http://www.msa.fr/lfr/sst/phyt-attitude

[13] https://osha.europa.eu/en/about-eu-osha

[14] SENSOR programme: https://www.cdc.gov/niosh/topics/pesticides/overview.html

5.3. Proposals for improvement of current framework of case incident reporting

In order to avoid duplication and waste of effort, a logical next step would be to now develop, with all concerned public and private sector actors, an EU 'phytopharmacovigilance' system for chemicals similar to the ones that have been put in place for medicines. This network could be based on committed and specifically trained occupational health physicians and general practitioners in rural areas, and resources should be allocated by Member States to establish and to successfully maintain the system. Indeed such a network would be useful in detecting acute effects; it would also act as a sentinel surveillance network for specific health effects (such as asthma, sensitisation, etc.) or for the detection of emerging work-related disease. In fact, while much experience has already been gained on how to gradually build such a system, it is nevertheless envisioned that this will take a number of years to be put in place. Several difficulties will arise because of the nature of the data collected (the sources of information are potentially diverse), the quality and completeness of the collected information for every case (especially the circumstances), the grading of severity and accountability of the observed effects (the link between the observed effect and the product). Rules should be defined so that they are identical from one 'evaluator' to another. The network should be stable over time (e.g. continuity in national organisations involved, consistent methodology employed, etc.), to ensure that the phytopharmacovigilance system fully complies with the objectives, i.e. monitoring changes over time. The use of phytopharmacovigilance data is unlikely to be limited to risk assessment purposes and may have an impact on risk management decisions (e.g. revisions in the terms and conditions of product authorisations or ultimately product withdrawal); this should be clear to all stakeholders from the outset.

Such a system may not merit being established solely for chemicals that are (predominantly) used as pesticides. However, given the legislative provisions already in place for pesticides, its development may need to be prioritised for pesticides.

In conclusion, the European Commission together with the Member States should initiate the development of an EU-wide vigilance framework for pesticides. These should include:

- harmonisation of human incident data collection activities at the EU level;
- coordination of the compilation of EU-wide databases;
- improving the collaboration between Poison Centres and regulatory authorities at national level in order to collect all the PPP poisonings produced in each Member State;
- guidance document on monitoring the impact of pesticide use on human health with harmonisation of data assessment for causal relationships;
- regular EU-wide reports.

6. Proposed use of epidemiological studies and vigilance data in support of the risk assessment of pesticides

This section briefly reviews the risk assessment process (Section 6.1) based on experimental studies and discusses what information epidemiological studies could add to that process. Next, the assessment of the reliability of epidemiological studies is addressed in Section 6.2. In Section 6.3, the relevance of one or more studies found to be reliable is assessed.

6.1. The risk assessment process

Risk assessment is the process of evaluating risks to humans and the environment from chemicals or other contaminants and agents that can adversely affect health. For regulatory purposes, the process used to inform risk managers consists of four steps (EFSA, 2012a). On the one hand, information is gathered on the nature of toxic effects (hazard identification) and the possible dose–response relationships between the pesticide and the toxic effects (hazard characterisation). On the other hand, information is sought about the potential exposure of humans (consumers, applicators, workers, bystanders and residents) and of the environment (exposure assessment). These two elements are weighed in the risk characterisation to estimate that populations be potentially exposed to quantities exceeding the reference dose values, that is, to estimate the extra risk of impaired health in the exposed populations. Classically, this is used to inform risk managers for regulatory purposes.

a) Step 1. *Hazard identification*.

Epidemiological studies and vigilance data are relevant for hazard identification as they can point to potential link between pesticide exposure and health. In this context, epidemiological data can provide invaluable information in 'scanning the horizon' for effects not picked up in experimental models. Importantly, these studies also provide information about potentially enhanced risks for vulnerable population subgroups, sensitive parts of the lifespan, and gender selective effects.

b) Step 2. *Hazard characterisation* (dose–response assessment). As previously discussed a classic dose–response framework is not normally considered when using epidemiological data as the exposure dose is rarely assigned. The challenge presented when high quality epidemiological studies are available is to see whether these can best be integrated into the scheme as numerical input. A dose–response framework is rarely considered when using epidemiological data for risk assessment of pesticides. However, previous scientific opinions of the EFSA CONTAM Panel have used epidemiology as basis for setting reference values, particularly in the case of cadmium, lead, arsenic and mercury, which are the most well-known and data rich (EFSA, 2009a,b, 2010b, 2012b). Even when they may not form the basis of a dose–response assessment, vigilance and epidemiological data may provide supportive evidence to validate or invalidate a dose–response study carried out in laboratory animals. Characterisation of the relationships between varying doses of a chemical and incidences of adverse effects in exposed populations requires characterisation of exposure or dose, assessment of response and selection of a dose–response model to fit the observed data in order to find a no-effect level. This raises two questions: can a dose–response be derived from epidemiological data to identify a no-effect level. If not, can epidemiological information otherwise contribute to the hazard characterisation?

Understanding dose–response relationships could also be relevant where adverse health outcomes are demonstrated to be associated with uses with higher exposures than EU good plant protection practice would give rise to, but where no association is observed from uses with lower exposures. It is clear that in this context the statistical summary of an epidemiological study defining RR or OR is potentially useful quantitative information to feed into the hazard characterisation process, when the study design meets the necessary standards.

c) Step 3. *Exposure assessment*. Data concerning the assessment of exposure are often hard to estimate in complex situations where a variety of uncontrolled 'real-world' factors confound the analysis. As discussed previously, contemporary biological monitoring is rarely carried out in the general human population for practical reasons including high cost, test availability and logistics. However, it is anticipated that in the near future biomonitoring studies and data on quantitative exposure to pesticides will increase.

Step 4. *Risk characterisation*. In this final step, data on exposure are compared with health-based reference values to estimate the extra risk of impaired health in the exposed populations. Human data can indeed help verify the validity of estimations made based on extrapolation from the full toxicological database regarding target organs, dose–response relationships and the reversibility of toxic effects, and to provide reassurance on the extrapolation process without direct effects on the definition of reference values (London et al., 2010).

Epidemiological data might also be considered in the context of uncertainty factors (UFs). An UF of 10 is generally used on animal data to account for interspecies variability of effects and this is combined with a further factor of 10 to account for variation in susceptibility of different parts of the human population. However, there are cases where only human data are considered (when this is more critical than animals data) and a single factor of 10 for intraspecies variability will apply. It is noted that at this moment Regulation (EC) No 1107/2009 Article 4(6) stipulates that: 'In relation to human health, no data collected on humans shall be used to lower the safety margins resulting from tests on animals'. The implication of this is that for risk assessment epidemiological data may only be used to increase the level of precaution used in the risk assessment, and not to decrease UFs even where relevant human data are available.

6.2. Assessment of the reliability of individual epidemiological studies

Factors to be considered in determining how epidemiology should be considered for a WoE assessment are described below and have been extensively outlined by available risk of bias tools for observational epidemiological studies.[15] The following examples represent factors to look for not an exhaustive list:

- *Study design and conduct*. Was the study design appropriate to account for the expected distributions of the exposure and outcome, and population at risk? Was the study conducted primarily in a hypothesis generating or a hypothesis-testing mode?

[15] Assessing Risk of Bias and Confounding in Observational Studies of Interventions or Exposures: Further Development of the RTI Item Bank (https://www.ncbi.nlm.nih.gov/books/NBK154464/) and Cochrane handbook.

- *Population*. Did the study sample the individuals of interest from a well-defined population? Did the study have adequate statistical power and precision to detect meaningful differences for outcomes between exposed and unexposed groups?
- *Exposure assessment*. Were the methods used for assessing exposure valid, reliable and adequate? Was a wide range of exposures examined? Was exposure assessed at quantitative level or in a categorical or dichotomous (e.g. ever vs never) manner? Was exposure assessed prospectively or retrospectively?
- *Outcome assessment*. Were the methods used for assessing outcomes valid, reliable and adequate? Was a standardised procedure used for collecting data on health outcomes? Were health outcomes ascertained independently from exposure status to avoid information bias?
- *Confounder control*: were potential confounding factors appropriately identified and considered? How were they controlled for? Were the methods used to document these factors valid, reliable and adequate?
- *Statistical analysis*. Did the study estimate quantitatively the independent effect of an exposure on a health outcome of interest? Were confounding factors appropriately controlled in the analyses of the data?
- Is the *reporting* of the study adequate and following the principles of transparency and the guidelines of the STROBE statement (or similar tools)?

Study evaluation should provide an indication on the nature of the potential limitations each specific study may have and an assessment of overall confidence in the epidemiological database.

Furthermore, the nature and the specificity of the outcome with regards to other known risk factors can influence the evaluation of human data for risk assessment purposes, particularly in case of complex health endpoints such as chronic effects with long induction and latency periods.

Table 2 shows the main parameters to be evaluated in single epidemiological studies and the associated weight (low, medium and high) for each parameter. Specific scientific considerations should be applied on a case-by-case basis, but it would be unrealistic to implement these criteria in a rigid and unambiguous manner.

Table 2: Study quality considerations for weighting epidemiological observational studies[(a),(b)]

Parameter	High	Moderate	Low
Study design and conduct	Prospective studies. Prespecified hypothesis (compound and outcome specific)	Case–control studies. Prospective studies not adequately covering exposure or outcome assessment	Cross-sectional, ecological studies Case–control studies not adequately covering exposure or outcome assessment
Population	Random sampling. Sample size large enough to warrant sufficient power Population characteristics well defined (including vulnerable subgroups)	Questionable study power, not justified in detail Non-representative sample of the target population Population characteristics not sufficiently defined	No detailed information on how the study population was selected Population characteristics poorly defined
Exposure assessment	Accurate and precise quantitative exposure assessment (human biomonitoring or external exposure) using validated methods Validated questionnaire and/or interview for chemical-specific exposure answered by subjects	Non-valid surrogate or biomarker in a specified matrix and external exposure Questionnaire and/or interview for chemical-specific exposure answered by subjects or proxy individuals	Poor surrogate Low-quality questionnaire and/or interview; information collected for groups of chemicals No chemical-specific exposure information collected; ever/never use of pesticides in general evaluated
Outcome Assessment	Valid and reliable outcome assessment. Standardised and validated in study population Medical record or diagnosis confirmed	Standardised outcome, not validated in population, or screening tool; or, medical record non-confirmed	Non-standardised and non-validated health outcome Inappropriate or self-reported outcomes.

Parameter	High	Moderate	Low
Confounder control	Adequate control for important confounders relevant to scientific question, and standard confounders Careful consideration is given to clearly indicated confounders	Confounders are partially controlled for Moderately control of confounders and standard variables Not all variables relevant for scientific question are considered	No control of potential confounders and effect modifiers in the design and analysis phases of the study
Statistical Analysis	Appropriate to study design, supported by adequate sample size, maximising use of data, reported well (not selective) Statistical methods to control for confounding are used and adjusted and unadjusted estimates are presented. Subgroups and interaction analysis are conducted	Acceptable methods, analytic choices that lose information, not reported clearly Post hoc analysis conducted but clearly indicated	Only descriptive statistics or questionable bivariate analysis is made Comparisons not performed or described clearly Deficiencies in analysis (e.g. multiple testing)
Reporting	Key elements of the Material and Methods, and results are reported with sufficient detail Numbers of individuals at each stage of study is reported A plausible mechanism for the association under investigation is provided	Some elements of the Material and Methods or results are not reported with sufficient detail Interpretation of results moderately addressed	Deficiencies in reporting (interpretation of effect estimates, confounder control) Selective reporting Paucity of information on relevant factors that may affect the exposure–health relationship. Misplaced focus of the inferential objectives Not justified conclusions

(a): Overall study quality ranking based on comprehensive assessment across the parameters.
(b): Adapted from US-EPA (2016), based in turn on Muñoz-Quezada et al. (2013) and LaKind et al. (2014).

If the above assessment is part of the evidence synthesis exercise, where epidemiological research is being assessed and quantitatively summarised, it permits more accurate estimation of absolute risk related to pesticide exposure and further quantitative risk assessment.

In the particular case of pesticide epidemiology data, three basic categories are proposed as a first tier to organise human data with respect to risk of bias and reliability[16]: (a) low risk of bias and high reliability (all or most of the above quality factors have been addressed with minor methodological limitations); (b) medium risk of bias and medium reliability (many of the above quality factors have been addressed with moderate methodological limitations); (c) high risk of bias and low reliability, because of serious methodological limitations or flaws that reduce the validity of results or make them largely uninterpretable for a potential causal association. The latter studies are considered unacceptable for risk assessment mainly because of poor exposure assessment, misclassification of exposure and/or health outcome, or lack of statistical adjustment for relevant confounders. Risk assessment should not be based on results of epidemiological studies that do not meet well-defined data quality standards. Furthermore, results of exploratory research will need to be confirmed in future research before they can be used for risk assessment.

6.3. Assessment of strength of evidence of epidemiological studies

This section briefly discusses some important issues specifically related to combining and summarising results from different epidemiological studies on the association between pesticides and human health.

The approach for weighting epidemiological studies is mainly based on the modified Bradford Hill criteria, which are a group of conditions that provide evidence bearing on a potentially causal relationship between an incidence and a possible consequence (strength, consistency, specificity, temporality, biological gradient, plausibility, coherence, experiment and analogy) (Table 3). Clearly, the

[16] These categories are in accordance with those currently used by EFSA for the peer review of pesticide active substances: acceptable, supplementary and unacceptable.

more of these criteria that are met the stronger the basis for invoking the association as evidence for a meaningful association. However, Bradford Hill was unwilling to define what causality was and never saw the criteria as sufficient or even absolutely necessary but simply of importance to consider in a common-sense evaluation.

Table 3: Considerations for WoE analysis based on the modified Bradford Hill criteria for evidence integration

Category	Considerations
Strength of Association	The assessment of the strength of association (not only the magnitude of association but also statistical significance) requires examination of underlying methods, comparison to the WoE in the literature and consideration of other contextual factors including the other criteria discussed herein
Consistency of Association	Associations should be consistent across multiple independent studies, particularly those conducted with different designs and in different populations under different circumstances. This criterion also applies to findings consistent across all lines of evidence (epidemiology, animal testing, *in vitro* systems, etc.) in light of modern data integration
Specificity	The original criteria of evidence linking a specific outcome to an exposure can provide a strong argument for causation has evolved and may have new and interesting implications within the context of data integration. Data integration may elucidate some mechanistic specificity among the varied outcomes associated with complex exposures. The lack of specificity can help to narrow down specific agents associated with disease
Temporality	Evidence of a temporal sequence between exposure to an agent and appearance of the effect within an appropriate time frame constitutes one of the best arguments in favour of causality. Thus, study designs that ensure a temporal progression between the two measures are more persuasive in causal inference
Biological Gradient (Dose–response)	Increased effects associated with greater exposures, or duration of exposures, strongly suggest a causal relationship. However, its absence does not preclude a causal association
Biological Plausibility	Data explained and supported by biologically plausible mechanisms based on experimental evidence strengthen the likelihood that an association is causal. However, lack of mechanistic data should not be taken as evidence against causality
Coherence	The interpretation of evidence should make sense and not to conflict with what is known about the biology of the outcome in question under the exposure-to-disease paradigm. If it does, the species closest to humans should be considered to have more relevance to humans
Experimental Evidence	Results from randomised experiments provide stronger evidence for a causal association than results based on other study designs. Alternatively, an association from a non-experimental study may be considered as causal if a randomised prevention derived from the association confirms the finding
Sequence of Key events	Provide a clear description of each of the key events (i.e. measurable parameters from a combination of *in vitro*, *in vivo* or human data sources) that underlie the established MoA/AOP for a particular health outcome. A fully elucidated MoA/AOP is a not requirement for using epidemiology studies in human health risk assessment

Adapted from Höfler (2005), Fedak et al. (2015) and US-EPA (2016).

For predictive causality, care must be taken to avoid the logical fallacy *post hoc ergo propter hoc* that states 'Since event Y followed event X, event Y must have been caused by event X'. Höfler (2005) quotes a more accurate 'counterfactual' definition as follows 'but for E, D will not occur or would not have occurred, but given E it will/would have occurred'. Yet, more detailed descriptions using symbolic logic are also available (Maldonado and Greenland, 2002). Rothman and Greenland (2008) stated that 'the only *sine qua non* for a counterfactual effect is the condition that the cause must precede the effect. If the event proposed as a result or "effect" precedes its cause, there may be an association between the events but certainly no causal relationship'.

6.3.1. Synthesis of epidemiological evidence

Systematic reviews and meta-analysis of observational studies can provide information that strengthens the understanding of the potential hazards of pesticides, exposure–response characterisation, exposure scenarios and methods for assessing exposure, and ultimately risk characterisation (van den Brandt, 2002). Systematic reviews entail a detailed and comprehensive plan

and search strategy defined *a priori* aimed at reducing bias by identifying, appraising and synthesising all relevant studies on a particular topic. The major steps of a systematic review are as follows: formulation of the research question; definition of inclusion and exclusion criteria; search strategy for studies across different databases; selection of studies according to predefined strategy; data extraction and creation of evidence tables; assessment of methodological quality of the selected studies; including the risk of bias; synthesis of data (a meta-analysis can be performed if studies allow); and interpretation of results and drawing of conclusions (EFSA, 2010a). Evidence synthesis is, however, challenging in the field of pesticide epidemiology as standardisation and harmonisation is difficult. Nonetheless, evidence synthesis should play a pivotal role in assessing the robustness and relevance of epidemiological studies.

Statistical tools have been developed that can help assess this evidence. When multiple studies on nearly identical sets of exposures and outcomes are available, these can provide important scientific evidence. Where exposure and outcomes are quantified and harmonised across studies, data from individual epidemiological studies with similar designs can be combined to gain enough power to obtain more precise risk estimates and to facilitate assessment of heterogeneity. Appropriate systematic reviews and quantitative synthesis of the evidence needs to be performed regularly (e.g. see World Cancer Research Fund approach to continuous update of meta-analysis for cancer risk factor[17]). Studies should be evaluated according to previously published criteria for observational research and carefully examine possible selection bias, measurement error, sampling error, heterogeneity, study design, and reporting and presentation of results.

Meta-analysis is the term generally used to indicate the collection of statistical methods for combining and contrasting the results reported by different studies (Greenland and O'Rourke, 2008). Meta-analysis techniques could be used to examine the presence of diverse biases in the field such as small study effects and excess significance bias. Meta-analyses, however, do not overcome the underlying biases that may be associated with each study design (i.e. confounding, recall bias or other sources of bias are not eliminated). The extent to which a systematic review or meta-analysis can draw conclusions about the effects of a pesticide depends strongly on whether the data and results from the included studies are valid, that is, on the quality of the studies considered. In particular, consistent findings among original studies resulting from a consistent bias will produce a biased conclusion in the systematic review. Likewise, a meta-analysis of invalid studies may produce a misleading result, yielding a narrow confidence interval around the wrong effect estimate.

In addition to summarising the basic study characteristics of the literature reviewed, a typical meta-analysis should include the following components: (a) the average effect size and effect size distribution for each outcome of interest and an examination of the heterogeneity in the effect size distributions; (b) subgroup analysis in which the variability present in the effect size distribution is systematically analysed to identify study characteristics that are associated with larger or smaller effect sizes; (c) publication bias analysis and other sensitivity analyses to assess the validity of conclusions drawn (Wilson and Tanner-Smith, 2014).

In a meta-analysis, it is important to specify a model that adequately describes the effect size distribution of the underlying population of studies. Meta-analysis using meaningful effect size distributions will help to integrate quantitative risk into risk assessment models. The conventional normal fixed- and random-effects models assume a normal effect size population distribution, conditionally on parameters and covariates. Such models may be adequate for estimating the overall effect size, but surely not for prediction if the effect size distribution exhibits a non-normal shape (Karabatsos et al., 2015).

6.3.2. Meta-analysis as a tool to explore heterogeneity across studies

When evaluating the findings of different studies, many aspects should be carefully evaluated. Researchers conducting meta-analyses may tend to limit the scope of their investigation to the determination of the size of association averaged over the considered studies. The motivation often is that aggregating the results yields greater statistical power and precision for the effect of interest. Because individual estimates of effect vary by chance, some variation is expected. However, estimates must be summarised only when meaningful. An important aspect that is often overlooked is heterogeneity of the strength of associations across subgroups of individuals. Heterogeneity between

[17] World Cancer Research Fund International. Continuous Update Project (CUP) http://www.wcrf.org/int/research-we-fund/continuous-update-project-cup

studies needs to be assessed and quantified when present (Higgins, 2008). In meta-analysis, heterogeneity among results from different studies may indeed be as informative as homogeneity. Exploring the reasons underlying any observed inconsistencies of findings is generally conducive of great understanding.

Figure 1 shows three forest plots from a fictitious example in which each of three pesticides (A, B and C) is evaluated in meta-analysis of two studies. It is assumed that both studies for each pesticide are of the highest quality and scientific rigor. No biases are suspected.

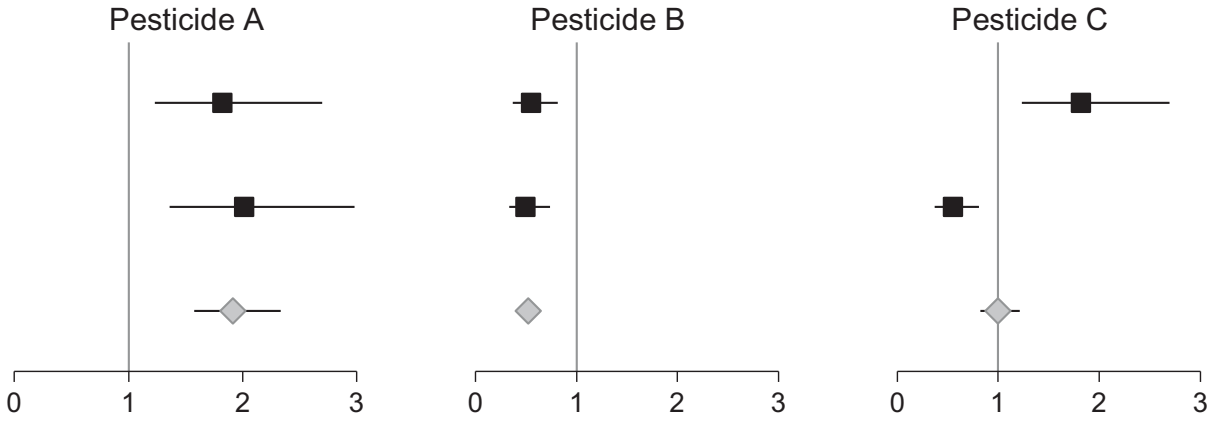


Figure 1: Forest plots from a fictitious example in which each of three pesticides (A, B and C) is evaluated in a meta-analysis of two studies. The x-axis in each plot represents the estimated risk ratio of the disease of interest comparing exposed and unexposed individuals. The squares denote the estimated risk ratio in each study and the grey diamonds the summarised risk ratio. The horizontal lines indicate 95% confidence intervals

The following text contains short comments on the interpretation of the results in Figure 1, one pesticide at a time.

- Exposure to pesticide A seems to double the risk of the disease. The results are consistent between the two studies and the confidence intervals do not contain the null value, one. These results, however, do not imply that (a) the risk ratio would be about 2 in any other study that was conducted on the same exposure and disease; or that (b) the risk ratio is two in any group of individuals (e.g. males or females, young or old).
- Exposure to pesticide B seems to halve the risk of the disease. The results are consistent between the two studies and the confidence intervals do not contain the null value, one. These results, however, do not imply that (a) the risk ratio would be about a half in any other study that was conducted on the same exposure and disease; or that (b) the risk ratio is about a half in any group of individuals (e.g. males or females, young or old).
- Exposure to pesticide C seems to double the risk of the disease in one study and to halve the risk in the other. The results are inconsistent between the two studies and the confidence intervals do not contain the null value, one. These results, however, do not imply that (a) the risk ratio would be about one in any other study that was conducted on the same exposure and disease; or that (b) the risk ratio is about one in any group of individuals (e.g. males or females, young or old).

What evidence can the results shown in Figure 1 provide?

The risk ratio reported by any study can be generalised to other populations only if all the relevant factors have been controlled for (Bottai, 2014; Santacatterina and Bottai, 2015). In this context, relevant factors are variables that are stochastically dependent with the health outcome of interest. For example, cardiovascular diseases are more prevalent among older subjects than among younger individuals. Age is therefore a relevant factor for cardiovascular diseases. The evidence provided by the results shown in Figure 1 are potentially valid only if this step was taken in each of the studies considered. If that was the case for the studies, then, there is evidence that exposure to pesticide A doubles the risk in the specific group of individuals considered by each of the two studies. If the risk ratios are summary measures over the respective study populations, then none of the findings should be generalised. However, if the risk ratios for pesticide A were not adjusted for any factor, and the underlying populations were very different

across the two studies, then there would still be evidence that there may be no relevant factors and pesticide A doubles the risk in any subgroup of individuals. Pesticide B appears to halve the risk, and the estimated confidence intervals are narrower for pesticide B than for pesticide A. Generalisability of the findings, however, holds for pesticide B under the conditions stated above for pesticide A. As for pesticide C, the forest plot provides evidence that exposure to this pesticide raises the risk of the disease in the group of individuals in one of the studies and decreases it in the group considered in the other study. Again, if the risk ratios are summary measures over the respective study populations, then none of the findings should be generalised. Investigating the reasons behind the inconsistency between the two studies on pesticide C can provide as much scientific insight as investigating the reasons behind the similarity between the studies on pesticide A or pesticide B.

In general, the overall summary measures provided by forest plots, such as the silver diamonds in each of the three panels of Figure 1, are of little scientific interest. When evaluating the findings of different studies, many aspects should be carefully evaluated. An important aspect that is often overlooked is heterogeneity of the strength of associations across subgroups of individuals. When information about subgroup analysis is provided in the publications that describe a study, this should be carefully evaluated. Sensitivity analyses should complement the results provided by different studies. These should aim to evaluate heterogeneity and the possible impact of uncontrolled for relevant factors along with information and sampling error. A synoptic diagram is displayed in Figure 2.

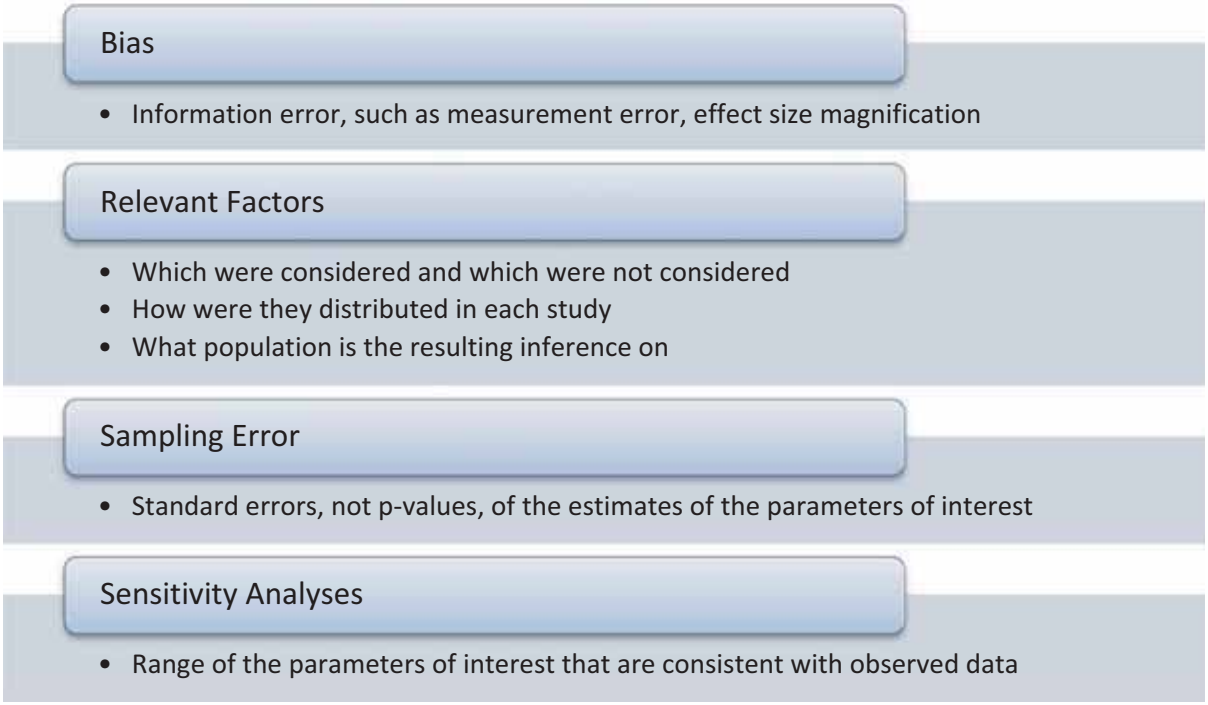


Figure 2: Items to consider when evaluating and comparing multiple studies

6.3.3. Usefulness of meta-analysis for hazard identification

Human data can be used for many stages of risk assessment. Single epidemiological studies, if further studies on the same pesticide are not available, should not be used as a sole source for hazard identification, unless they are high quality studies (according to criteria shown in Table 2). Evidence synthesis techniques which bring together many studies, such as systematic reviews and meta-analysis (where appropriate) should be utilised instead. Although many meta-analyses have been carried out for the quantitative synthesis of data related to chronic diseases, their application for risk assessment modelling is still limited.

Importantly, evidence synthesis will provide a methodological assessment and a risk of bias assessment of the current evidence highlighting areas of uncertainties and identifying associations with robust and credible evidence.

Figure 3 shows a simple methodology proposed for the application of epidemiological studies into risk assessment. The first consideration is the need of combining different epidemiological studies

addressing the same outcome. This can be made following criteria proposed by EFSA guidance for systematic reviews (EFSA, 2010a). Then, the risk of bias is assessed based on the factors described in Section 6.2 for a WoE assessment, namely: study design and conduct, population, exposure assessment, outcome assessment, confounder control, statistical analysis and reporting of results. Those studies categorised as of low reliability will be considered unacceptable for risk assessment. The remaining studies will be weighted and used for hazard identification.

If quantitative data are available, a meta-analysis can be conducted to create summary data and to improve the statistical power and precision of risk estimates (OR, RR) by combining the results of all individual studies available or meeting the selection criteria. As meta-analyses determine the size of association averaged over the considered studies, they provide a stronger basis for hazard identification. Moreover, under certain circumstances, there is the possibility to move towards risk characterisation metrics because these measured differences in health outcomes (OR, RR) can be converted to dose–response relationships (Nachman et al., 2011). Although quite unusual in practice, this would allow for the identification of critical effects in humans and/or setting reference values without the need of using animal extrapolation.

Since heterogeneity is common in meta-analyses, there is a need to assess which studies could be combined quantitatively. Heterogeneity can be genuine, representing diverse effects in different subgroups, or might represent the presence of bias. If heterogeneity is high (I^2 greater than 50%), individual studies should not be combined to obtain a summary measure because of the high risk of aggregating bias from different sources. Sources of heterogeneity should be explored through sensitivity analysis and/or meta-regression. Furthermore, the presence of diverse biases in the meta-analysis should be examined, such as small study effects, publication bias and excess significance bias. It is important to find models that adequately describe the effect size distribution of the underlying studied populations.

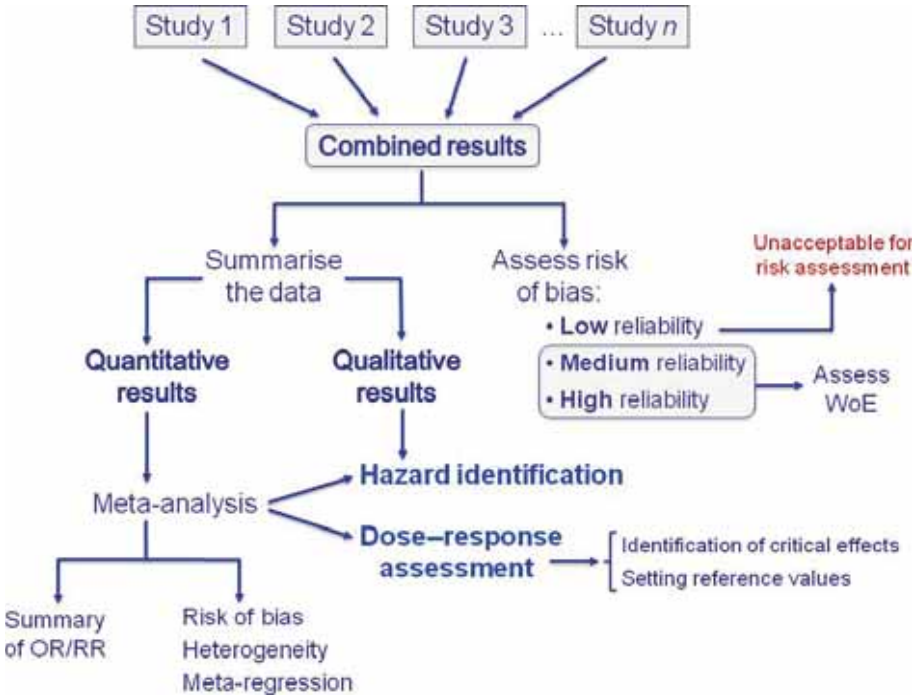


Figure 3: Methodology for utilisation of epidemiological studies for risk assessment

6.3.4. Pooling data from similar epidemiological studies for potential dose–response modelling

As in other fields of research, findings from a single epidemiological study merit verification through replication. When the number of replications is abundant, it may be worthwhile to assess the entire set of replicate epidemiological studies through a meta-analysis and ascertain whether, for key outcomes, findings are consistent across studies. Such an approach will provide more robust conclusions about the existence of cause-effect relationships.

Once a hazard has been identified, the next step in risk assessment is to conduct a dose–response assessment to estimate the risk of the adverse effect at different levels of exposure and/or the concentration level below which no appreciable adverse health effect can be assumed for a given population. However, this step requires fully quantitative (or at least semi-quantitative) exposure data at an individual level. Summary estimates resulting from quantitative synthesis would be more informative for risk assessment if they present an OR for a given change in the continuous variable of exposure (or per a given percentile change in exposure) as this allows for relative comparisons across studies and could be of help to derive health-based reference values. Only within such a framework can data from human studies with similar designs be merged to gain enough power to model proper dose–response curves (Greenland and Longnecker, 1992; Orsini et al., 2012).

Conversely, meta-analytical approaches may be of limited value if a combined OR is calculated based on meta-analyses interpreting exposure as a 'yes' or a 'no' (ever vs never) because exposures are not necessarily to active ingredients in the same proportion in all studies included. Even though in these cases, meta-analyses may consistently find an increased risk associated with pesticide exposure, for risk assessment the exposure needs to characterise the effect of specific pesticide classes or even better individual pesticides as their potency may differ within the same class (Hernández et al., 2016).

This approach would allow points of departure to be identified (e.g. benchmark doses (BMD)) and would be relevant for the integration of epidemiological studies into quantitative risk assessment. Although BMD modelling is currently used for analysing dose–response data from experimental studies, it is possible to apply the same approach to data from observational epidemiological studies (Budtz-Jørgenson et al., 2004). The EFSA Scientific Committee confirmed that the BMD approach is a scientifically more advanced method compared to the no observed-adverse-effect level (NOAEL) approach for deriving a Reference Point, since it makes extended use of the dose–response data from experimental and epidemiological studies to better characterise and quantify potential risks. This approach, in principle, can be applicable to human data (EFSA Scientific Committee, 2017b), although the corresponding guidelines are yet to be developed.

Dose–response data from observational epidemiological studies may differ from typical animal toxicity data in several respects and these differences are relevant to BMD calculations. Exposure data often do not fall into a small number of well-defined dosage groups. Unlike most experimental studies, observational studies may not include a fully unexposed control group, because all individuals may be exposed to some extent to a chemical contaminant. In this case, the BMD approach still applies since fitting a dose–response curve does not necessarily require observations at zero exposure. However, the response at zero exposure would then need to be estimated by low-dose extrapolation. Hence, the BMD derived from epidemiological data can be strongly model-dependent (Budtz-Jørgensen et al., 2001).

Epidemiology data need to be of sufficient quality to allow the application of the BMD approach, especially in terms of assigning an effect to a specific pesticide and its exposure. Clear rules and guidance, and definition of model parameters need to be considered for such a BMD approach, which might differ from BMD approaches from controlled experimental environments. Although the BMD modelling approach has been applied to epidemiological data on heavy metals and alcohol (Lachenmeier et al., 2011), currently, few individual studies on pesticides are suitable for use in dose–response modelling, much less in combination with other studies. However, future studies should be conducted and similarly reported so that they could be pooled together for a more robust assessment.

7. Integrating the diverse streams of evidence: human (epidemiology and vigilance data) and experimental information

This section first considers in Section 7.1 the different nature of the main streams of evidence, i.e. originating either from experimental studies or from epidemiological studies. The approach used is that recommended by the EFSA Scientific Committee Guidance on WoE (EFSA Scientific Committee, 2017b), which distinguishes three successive phases to assess and integrate these different streams of information: reliability, relevance and consistency. The first step, consists in the assessment of the reliability of individual studies be they epidemiological (addressed in Section 6) or experimental (beyond the scope of this Scientific Opinion). Then, the relevance (strength of evidence) of one or more studies found to be reliable is assessed using principles of epidemiology (addressed in Section 6) and toxicology. Next, Section 7.2 considers how to bring together different streams of relevant information from epidemiological and experimental studies, which is considered in a WoE approach, to assess consistency and biological plausibility for humans.

7.1. Sources and nature of the different streams of evidence Comparison of experimental and epidemiological approaches

In the regulatory risk assessment of pesticides, the information on the toxic effects is based on the results of a full set of experiments as required by Regulation (EC) 283/2013 and 284/2013, and conducted according to OECD guidelines. They are carried out *in vivo* or *in vitro*, so there will always be some high-quality experimental data available for pesticides as required to be provided by applicants under Regulation (EC) 1107/2009. A number of categories are established for rating the reliability of each stream of evidence according to the EFSA peer review of active substances: acceptable, supplementary and unacceptable. The data quality and reliability of *in vivo* or *in vitro* toxicity studies should be assessed using evaluation methods that better provide more structured support for determining a study's adequacy for hazard and risk assessments. Criteria have been proposed for conducting and reporting experimental studies to enable their use in health risk assessment for pesticides (Kaltenhäuser et al., 2017).

Animal (*in vivo*) studies on pesticide active substances conducted according to standardised test guidelines and good laboratory practices (GLP, e.g. OECD test guidelines) are usually attributed higher reliability than other research studies. Notwithstanding, since there is no evidence that studies conducted under such framework have a lower risk of bias (Vandenberg et al., 2016), evidence from all relevant studies, both GLP and non-GLP, should also be considered and weighted. Thus, data from peer-reviewed scientific literature should be taken into account for regulatory risk assessment of pesticide active substances, provide they are of sufficient quality after being assessed for methodological reliability. Their contribution to the overall WoE is influenced by factors including test organism, study design and statistical methods, as well as test item identification, documentation and reporting of results (Kaltenhäuser et al., 2017).

The internal validity of *in vitro* toxicity studies should be evaluated as well to provide a better support for determining a study's adequacy for hazard and risk assessments. *In silico* modelling can be used to derive structure–activity relationships (SAR) and to complement current toxicity tests for the identification and characterisation of the mode or mechanisms of action of the active substance in humans. These alternative toxicity testing (and non-testing) approaches could be helpful in the absence of animal data, e.g. to screen for potential neurodevelopmental or endocrine disruption effects of pesticides, and to increase confidence in animal testing. Considering the demand for minimising the number of animal studies for regulatory purposes, non-animal testing information can provide relevant stand-alone evidence that can be used in the WoE assessment.

A number of toxicological issues are amenable for systematic review, from the impact of chemicals on human health to risks associated with a specific exposure, the toxicity of chemical mixtures, the relevance of biomarkers of toxic response or the assessment of new toxicological test methods (Hoffmann et al., 2017). For instance, in a previous Scientific Opinion EFSA used a systematic review for the determination of toxicological mechanisms in the frame of AOP approach (Choi et al., 2016; EFSA Scientific Committee, 2017c).

Besides toxicity data on the active substance, such data may also be required on metabolites or residues if human exposure occur through the diet or drinking water. Results from these studies are then considered in relation to expected human exposures estimated through food consumption and other sources of exposure. The strength of this approach is that *in vivo* studies account for potential toxic metabolites, though not always animal metabolic pathways parallels the ones of humans.

Experimental studies in laboratory animals are controlled studies where confounding is eliminated by design, which is not always the case with epidemiological studies. Animals used in regulatory studies are, however, typically inbred, genetically homogeneous and due to the controlled environment they lack the full range of quantitative and qualitative chemical susceptibility profiles. Nevertheless, animal surrogates of human diseases are being challenged by their scientific validity and translatability to humans, and the lack of correlation often found between animal data and human outcomes can be attributed to the substantial interspecies differences in disease pathways and disease-induced changes in gene expression profiles (Esch et al., 2015). Thereby, many experimental models do not capture complex multifactorial diseases making animal-to-human extrapolation subject to considerable uncertainty. Current risk assessment is therefore by its nature predictive and may be insufficient because it is chemical-specific and humans are exposed to a large number of chemicals from environmental, dietary and occupational sources or because of different toxicokinetic differences. In recognition of the uncertain nature of animal-to-human extrapolation, the regulatory risk assessment advice does not just consider the relevant point(s) of departure (NOAEL, LOAEL or BMDL) that have

been identified as safe but lowers these values using uncertainty factors (UFs) to propose safe reference dose values, either for acute or chronic toxicity.

Given the limitations of studies in laboratory animals, epidemiological studies in the 'real world' are needed, even if they have limitations of their own. Epidemiological studies incorporate the true (or estimated) range of population exposures, which usually are intermittent and at inconsistent doses instead of occurring at a consistent rate and dose magnitude (Nachman et al., 2011). Since epidemiological studies are based on real-world exposures, they provide insight into actual human exposures that can then be linked to diseases, avoiding the uncertainty associated with extrapolation across species. Hence, it can be said that they address the requirements of Regulation 1107/2009 Article 4, which stipulates that the risk assessment should be based on good plant protection practice and realistic use conditions. Thus, epidemiological studies assist problem formulation and hazard/risk characterisation whilst avoiding the need for high dose extrapolation (US-EPA, 2010).

Epidemiological studies therefore provide the opportunity to (a) identify links with specific human health outcomes that are difficult to detect in animal models; (b) affirmation of the human relevance of effects identified in animal models; (c) ability to evaluate health effects for which animal models are unavailable or limited (Raffaele et al., 2011). Epidemiological evidence will be considered over experimental animal evidence only when sufficiently robust pesticide epidemiological studies are available. However, in epidemiological studies, there are always a variety of factors that may affect the health outcome and confound the results. For example, when epidemiological data suggest that exposures to pesticide formulations are harmful they usually cannot identify what component may be responsible due to the complexity of accurately assessing human exposures to pesticides. While some co-formulants are not intrinsically toxic, they can be toxicologically relevant if they change the toxicokinetics of the active substance. In addition, confounding by unmeasured factor(s) associated with the exposure can never be fully excluded; however, a hypothetical confounder (yet unrecognised) may not be an actual confounder and has to be strongly associated with disease and exposure in order to have a meaningful effect on the risk (or effect size) estimate, which is not always the case.

Many diseases are known to be associated with multiple risk factors; however, a hazard-by-hazard approach is usually considered for evaluating the consequences of individual pesticide hazards on vulnerable systems (Figure 4A). Specifically, single-risk analysis allows a determination of the individual risk arising from one particular hazard and process occurring under specific conditions, while it does not provide an integrated assessment of multiple risks triggered by different environmental stressors (either natural or anthropogenic) (Figure 4B). Risk assessment would benefit by developing procedures for evaluating evidence for co-occurrence of multiple adverse outcomes (Nachman et al., 2011), which is more in line with what happens in human setting. For these reasons, if appropriately conducted, epidemiological studies can be highly relevant for the risk assessment process.

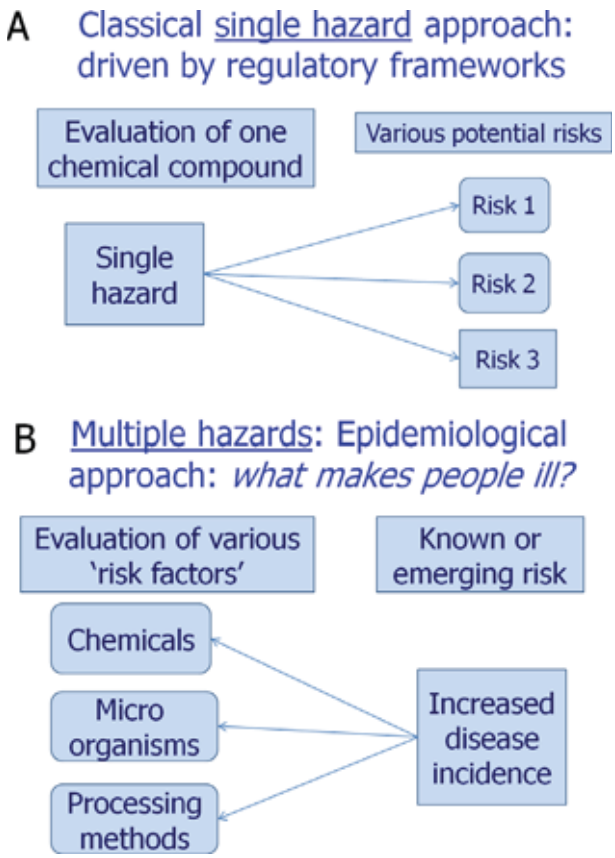


Figure 4: Role of epidemiological studies when compared to classical toxicological studies

In parallel with epidemiological data, vigilance data can provide an additional stream of evidence, especially for acute toxicity. Cases are usually well-documented and information can be used at different steps of the risk assessment; these include: level and duration of exposure, clinical course and assessment of the causal relationship. In severe cases, the toxin and/or the metabolites are usually measured in blood or urine which allows for comparison with animal data and in some cases for setting toxicological values.

In summary, experimental studies or epidemiological studies and vigilance data represent two different approaches to collect and assess evidence i.e. one emanating from controlled exposures (usually to a single substance) using experimental study design and a relatively homogeneous surrogate population, the other reflecting the changes observed in a heterogeneous target population from mixed (and varying) exposure conditions using non-experimental study design (ECETOC, 2009). Epidemiology and toxicology each bring important and different contributions to the identification of human hazards. This makes both streams of evidence complementary, and their combination represents a powerful approach. Animal studies should always inform the interpretation of epidemiological studies and vice versa; hence, they should not be studied and interpreted independently.

7.2. Principles for weighting of human observational and laboratory animal experimental data

Following the identification of reliable human (epidemiological or vigilance) studies and the assessment of the relevance of the pooled human studies, the separate lines of evidence that were found to be relevant need to be integrated with other lines of evidence that were equally found to be relevant.

The first consideration is thus how well the health outcome under consideration is covered by toxicological and epidemiological studies. When both animal and human studies are considered to be available for a given outcome/endpoint, this means that individual studies will first have been assessed for reliability and strength of evidence (Sections 6.2 and 6.3, respectively, for epidemiological studies)

prior to the weighting of the various sources of evidence. Although the different sets of data can be complementary and confirmatory, individually they may be insufficient and pose challenges for characterising properly human health risks. Where good observational data are lacking, experimental data have to be used. Conversely, when no experimental data is available, or the existing experimental data were found not to be relevant to humans, the risk assessment may have to rely on the available and adequate observational studies.

A framework is proposed for a systematic integration of data from multiple lines of evidence (in particular, human and experimental studies) for risk assessment (Figure 5). Such integration is based on a WoE analysis accounting for relevance, consistency and biological plausibility using modified Bradford Hill criteria (Table 3). For a comparative interpretation of human and animal data, this framework should rely on the following principles (adapted from ECETOC, 2009; Lavelle et al., 2012):

- Although the totality of evidence should be assessed, only the studies that are found to be reliable (those categorised as acceptable or supplementary evidence) are considered further. If the data from the human or the experimental studies is considered to be of low reliability (categorised as unacceptable), no risk assessment can be conducted.
- A WoE approach should be followed where several lines of evidence are found to be relevant. For pesticide active substances, experimental studies following OECD test guidelines are deemed high reliability unless there is evidence to the contrary. The strength of evidence from animal studies can be upgraded if there is high confidence in alternative pesticide toxicity testing or non-testing methods (e.g. *in vitro* and *in silico* studies, respectively). As for epidemiological evidence, the conduct of meta-analysis provides a more precise estimate of the magnitude of the effect than individual studies and also allows for examining variability across studies (see Section 6.3).
- Next, the studies that are found to be more relevant for the stage being assessed are to be given more weight, regardless of whether the data comes from human or animal studies. Where human data are of highest relevance, and supported by a mechanistic scientific foundation, they should take precedence for each stage of the risk assessment. When human and experimental data are of equal or similar relevance, it is important to assess their concordance (consistency across the lines of evidence) in order to determine whether and which data set may be given precedence.
    - In case of concordance between human and experimental data, the risk assessment should use all the data as both yield similar results in either hazard identification (e.g. both indicate the same hazard) or hazard characterisation (e.g. both suggest similar safe dose levels). Thus, both can reinforce each other and similar mechanisms may be assumed in both cases.
    - In case of non-concordance, the framework needs to account for this uncertainty. For hazard identification, the data suggesting the presence of a hazard should generally take precedence. For dose–response, the data resulting in the lower acceptable level should take precedence. In every situation of discordance, the reasons for this difference should be considered. If the reason is related to the underlying biological mechanisms, or toxicokinetic differences between humans and animal models, then confidence in the risk assessment will increase. Conversely, if the reason cannot be understood or explained, then the risk assessment may be less certain. In such cases, efforts should be made to develop a better understanding of the biological basis for the contradiction.

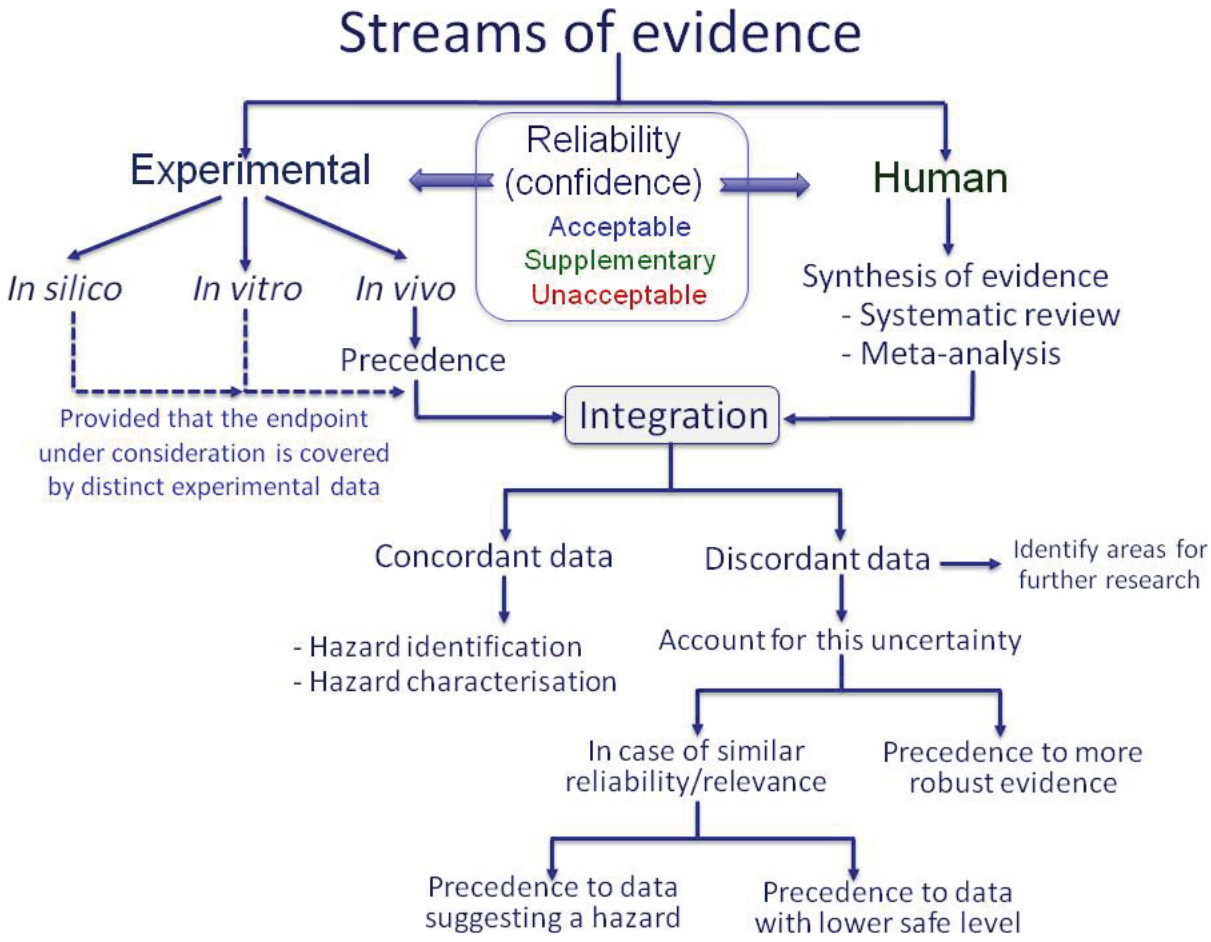


Figure 5: Methodology for the integration of human and animal data for risk assessment

Epidemiological studies provide complementary data to analyse risk and should be contextualised in conjunction with well-designed toxicological *in vivo* studies and mechanistic studies. The overall strength of the evidence achieved from integrating multiple lines of evidence will be at least as high as the highest evidence obtained for any single line. This integrated approach provides explicit guidance on how to weight and integrate toxicological and epidemiological evidence. This is a complex task that becomes even more difficult when epidemiological data deal with multifactorial, multihit, chronic diseases for which toxicological models, or disease-specific animal models, are limited.

7.3. Weighting all the different sources of evidence

The WHO/IPCS defines the WoE approach as a process in which all of the evidence considered relevant for risk assessment is evaluated and weighted (WHO/IPCS, 2009). The WoE approach, taking the risk assessment of chemical substances as an example, requires the evaluation of distinct lines of evidence (*in vivo*, *in vitro*, *in silico*, population studies, modelled and measured exposure data, etc.). The challenge is to weight these types of evidence in a systematic, consistent and transparent way (SCENIHR, 2012). The weighting may be formally quantitative or rely on categorisation according to criterion referencing of risk.

An EFSA Working Group was established to provide transparent criteria for the use of the WoE approach for the evaluation of scientific data by EFSA's Panels and Scientific Committee (EFSA, 2015b). The aim of this Working Group was to provide support to stakeholders on how individual studies should be selected and weighted, how the findings integrated to reach the final conclusions and to identify uncertainties regarding the conclusions.

The WoE approach is not consistently considered in the risk assessment of pesticides in the peer review process of DAR or RAR. Expert judgement alone, without a structured WoE approach, has been more commonly used. A few examples can be found, such as the peer review of glyphosate (EFSA, 2015c), where the rapporteur Member State (RMS) considered all the data either from industry or

from public literature, including epidemiological data, and took a specific WoE approach with established *ad hoc* criteria and considering all data available for proposing an 'overall' NOAEL for each endpoint of toxicity explored.

The US-EPA has recently applied specific criteria for the WoE approach to the peer review of the pesticide chlorpyrifos by following the 'Framework for incorporating human epidemiologic & incident data in health risk assessment'. In this specific case, a WoE analysis has been conducted to integrate quantitative and qualitative findings across many lines of evidence including experimental toxicology studies, epidemiological studies and physiologically based pharmacokinetic and pharmacodynamic (PBPK-PD) modelling. Chlorpyrifos was also used as an example for the EFSA Guidance on literature search under Regulation (EC) No 1107/2009. In addition, an EFSA conclusion (EFSA, 2014a) took into consideration the US-EPA review (2011) to revise its first conclusion produced in 2011.

In sum, a broader WoE approach can be applied to evaluate the available scientific data using modified Bradford Hill criteria as an organisational tool to increase the likelihood of an underlying causal relationship (Table 3). Although epidemiology increasingly contributes to establishing causation, an important step to this end is the establishment of biological plausibility (US-EPA, 2010; Adami et al., 2011; Buonsante et al., 2014).

7.4. Biological mechanisms underlying the outcomes

A biological mechanism describes the major steps leading to a health effect following interaction of a pesticide with its biological targets. The mechanism of toxicity is described as the major steps leading to an adverse health effect. An understanding of all steps leading to an effect is not necessary, but identification of the key events following chemical interaction is required to describe a mechanism (of toxicity in the case of an adverse health effect). While many epidemiological studies have shown associations between pesticide exposures and chronic diseases, complementary experimental research is needed to provide mechanistic support and biological plausibility to the human epidemiological observations. Experimental exposures should be relevant to the human population provided that the biologic mechanisms in laboratory animals occur in humans.

Establishing biological plausibility as part of the interpretation of epidemiological studies is relevant and should take advantage of modern technologies and approaches (Section 7.6). In this context, the AOP framework can be used as a tool for systematically organising and integrating complex information from different sources to investigate the biological mechanisms underlying toxic outcomes and to inform the causal nature of links observed in both experimental and observational studies (Section 7.5).

The use of data to inform specific underlying biological mechanisms or pathways of the potential toxic action of pesticides is limited since only selected pesticide chemicals have been investigated for biological function in relation to a specific health outcome. It may be possible to formulate a mode of action (MoA) hypothesis, particularly where there is concordance between results of comparable animal studies or when different chemicals show the same pattern of toxicity. It is essential to identify the toxicant and the target organ as well as the dose–response curve of the considered effect and its temporal relationship. If the different key events leading to toxicity and a MoA hypothesis can be identified, it is sometimes possible to evaluate the plausibility of these events to humans (ECETOC, 2009).

Sulfoxaflor is an example where MoA has been extensively studied and has been also widely used as an example during the ECHA/EFSA MOA/HRF workshop held in November 2014. Sulfoxaflor induced hepatic carcinogenicity in both rats and mice. Studies to determine the MoA for these liver tumours were performed in an integrated and prospective manner as part of the standard battery of toxicology studies such that the MoA data were available prior to, or by the time of, the completion of the carcinogenicity studies. The MoA data evaluated in a WoE approach indicated that the identified rodent liver tumour MoA for sulfoxaflor would not occur in humans. For this reason, sulfoxaflor is considered not to be a potential human liver carcinogen.

Furthermore, sometimes MoA data may indicate a lack of possible effects. If there are biological data that indicate an adverse effect is not likely to occur in humans, this should inform the interpretation of epidemiological studies. Nevertheless, while primary target site selectivity between pests and humans plays an important role in pesticides safety, secondary targets in mammals must also be considered.

In the case of exposure to multiple pesticides, the decision to combine risks can be taken if the pesticides share a common mechanism of toxicity (act on the same molecular target at the same target tissue, act by the same biochemical mechanism of action, and share a common toxic intermediate) which may cause the same critical effect or just based on the observation that they share the same target organ (EFSA 2013a,b). However, cumulative risk assessment is beyond the scope of this Opinion.

7.5. Adverse Outcome Pathways (AOPs)

The AOP methodology provides a framework to collect and evaluate relevant chemical, biological and toxicological information in such a way that is useful for risk assessment (OECD, 2013). An AOP may be defined as the sequence of key events following the interaction of a chemical with a biological target (molecular initiating event (MIE)) to the *in vivo* adverse outcome relevant to human health. All these key events are necessary elements of the MoA and should be empirically observable or constitute biologically based markers for such an event. An AOP is therefore a linear pathway from one MIE to one adverse outcome at a level of biological organisation relevant to risk assessment. The goal of an AOP is to provide a flexible framework to describe the cascade of key events that lead from a MIE to an adverse outcome in a causal linkage (EFSA PPR Panel, 2017). The 'key events' must be experimentally measurable and the final adverse effect is usually associated with an *in vivo* OECD Test Guideline. However, in some cases the adverse outcome may be at a level of biological organisation below that of the apical endpoint described in a test guideline (OECD, 2013).

A particular MIE may lead to several final adverse effects and, conversely, several MIEs may converge in the same final adverse effect. However, each AOP will have only one MIE and one final adverse effect, but may involve an unlimited number of intermediate steps (Vinken, 2013). It should be noted that key events at different levels of biological organisation provide a greater WoE than multiple events at the same level of organisation (OECD, 2013).

The essential biochemical steps involved in a toxic response are identified and retrieved from an in-depth survey of relevant scientific literature or from experimental studies. Any type of information can be incorporated into an AOP, including structural data, 'omics-based' data and *in vitro*, *in vivo* or *in silico* data. However, *in vivo* data are preferred over *in vitro* data and endpoints of interest are preferred to surrogate endpoints (Vinken, 2013). The AOPs identified must not be incompatible with normal biological processes, since they need to be biologically plausible.

Qualitative AOPs (intended as an AOP including the assembly and evaluation of the supporting WoE following the OECD guidance for AOP development) should be the starting and standard approach in the process of integration of epidemiology studies into risk assessment by supporting (or identifying the lack of support for) the biological plausibility of the link between exposure to pesticides affecting the pathway and the adverse outcome. Accordingly, qualitative AOPs may be developed solely for the purpose of hazard identification, to support biological plausibility of epidemiological studies based on mechanistic knowledge (EFSA PPR Panel, 2017).

The AOP framework is a flexible and transparent tool for the review, organisation and interpretation of complex information gathered from different sources. This approach has the additional advantage of qualitatively characterising the uncertainty associated with any inference of causality and identifying whether additional mechanistic studies or epidemiological research would be more effective in reducing uncertainty. The AOP framework is therefore a useful tool for risk assessment to explore whether an adverse outcome is biologically plausible or not. For the purpose of analysing the biological plausibility, AOPs can serve as an important tool, particularly when the regulatory animal toxicological studies are negative but the evaluation of the apical endpoint (or relevant biomarkers) observed in epidemiological studies is considered inadequate based on the AOP. By means of mechanistically describing apical endpoints, the AOP contributes to the hazard identification and characterisation steps in risk assessment. As the AOP framework is chemically agnostic, if complemented by the MoA and/or Integrated Approach on Testing and Assessment (IATA) framework, it will support the chemical specific risk assessment (EFSA PPR Panel, 2017).

AOP and MoA data can be used to assess the findings of epidemiological studies to weight their conclusions. Whether those findings are inconsistent with deep understanding of biological mechanisms, or simply empirical, they should be given less weight than other findings that are consistent with AOP or MoA frameworks once established. However, there are relatively few examples of well-documented AOPs and a full AOP/MoA framework is not a requirement for using epidemiological studies in risk assessment.

AOPs are thus a critical element to facilitate moving towards a mechanistic-based risk assessment instead of the current testing paradigm relying heavily on apical effects observed in animal studies. Shifting the risk assessment paradigm towards mechanistic understanding would reduce limitations of the animal data in predicting human health effects for a single pesticide, and also support the current efforts being made on cumulative risk assessment of pesticide exposure (EFSA PPR Panel, 2017).

7.6. Novel tools for identifying biological pathways and mechanisms underlying toxicity

The elucidation of toxicity pathways brings the opportunity of identifying novel biomarkers of early biological perturbations in the toxicodynamic progression towards overt disease, particularly from advances in biomonitoring, in -omics technologies and systems biology (toxicology). The revolution of omics in epidemiology holds the promise of novel biomarkers of early effect and offers an opportunity to investigate mechanisms, biochemical pathways and causality of associations.

The growing recognition of the value of biomonitoring data in epidemiological investigations may help to reduce misclassification by providing objective measures of exposure and outcome. As long as biomarker data for exposure, outcome and susceptibility are increasingly generated, epidemiology will have a greater impact in the understanding of toxicodynamic progression as a function of pesticide exposure and eventually in risk assessment. A challenge for risk assessors will be to acknowledge where subtle and early changes along the toxicodynamic pathway are indicative of increased potential for downstream effects (Nachman et al., 2011). Omics data can be used for gaining insight to the MoA by identifying pathways affected by pesticides and as such can assist hazard identification, the first step in risk assessment.

Transcriptomic, metabolomic, epigenomic and proteomic profiles of biological samples provide a detailed picture, sometimes at individual molecule resolution, of the evolving state of cells under the influence of environmental chemicals, thus revealing early mechanistic links with potential health effects. Nowadays, the challenges and benefits that advances in -omics techniques can bring to regulatory toxicology are still being explored (Marx-Stoelting et al., 2015). Clear rules for assessing the specificity of these biomarkers are necessary.

Those -omic applications most relevant and advanced in the context of toxicology are analysis of MoA and the derivations of AOP, and biomarker identification, all of which potentially assist epidemiology too. For example, (a) transcriptomics: comparing gene expression (mRNA) profiles can be used for biomarker discovery, grouping expressed genes into functional groups (Gene Ontology categories) or for Gene Set Analysis. Such techniques may provide varying information regarding biological mechanisms. (b) Proteomics: studying the protein profile of samples, with sophisticated analysis of protein quantity and post-translational modifications which may be associated with changes in biological pathways following exposure and possible disease development, utilising informatics and protein databases for identification and quantification. (c) Metabolomics uses nuclear magnetic resonance spectroscopy or mass-spectrometry based techniques to produce data which are analysed via software, and databases, to identify markers (molecular signatures and pathways) that correlate with exposure or disease. (d) The use of the exposome (the totality of exposures received by an individual during life) might be better defined by using -omics technologies and biomarkers appropriate for human biomonitoring. Nevertheless, important limitations stemming from the lack of validation of these methodologies and their cost limit their use at large scale.

The application of -omics technologies to environmental health research requires special consideration to study design, validation, replications, temporal variance and meta-data analysis (Vlaanderen et al., 2010). For larger studies, intra-individual variability in the molecular profiles measured in biological samples should show less variability than the interindividual variation in profiles of gene expression, protein levels or metabolites, which are highly variable over time. It is important that these inter-individual variations should not be larger than variation related to exposure changes, but it is not certain if this will be true.

The biologically meaningful omics signatures identified by performing omics-exposure and omics-health association studies provide useful data for advanced risk assessment. This approach supports moving away from apical toxicity endpoints towards earlier key events in the toxicity pathway resulting from chemical-induced perturbation of molecular/cellular responses (NRC, 2007).

7.7. New data opportunities in epidemiology

The current technological landscape permits the digitisation and storage of unprecedented amount of data from many sources, including smart phones, text messages, credit card purchases, online activity, electronic medical records, global positioning system (GPS) and supermarket purchasing data. While some of these data sources may provide valuable information for risk assessment, many of them contain personal information that can outpace legal frameworks and arise questions about the ethics of its use for scientific or regulatory purposes. A specific example is constituted by data containing

personal information related to health, which are considered sensitive or especially protected, such as electronic medical records, information from occupational or environmental questionnaires, geographic location, health or social security number, etc. These various forms of health information are being easily created, stored and accessed. Big data provide researchers with the ability to match or link records across a number of data sources. Linking of big data sources of health and heritable information offers great promise for understanding disease predictors (Salerno et al., 2017); however, there are challenges in using current methods to process, analyse and interpret the data systematically and efficiently or to find relevant signals in potential oceans of noise, as noted by the Board on Environmental Studies and Toxicology of the National Academies of Sciences, Engineering, and Medicine in its 2017 report.[18]

In addition, medico-administrative data, such as drug reimbursements drawn from National Health Insurance or hospital discharge databases, can be cross-linked with data on agricultural activities drawn from agricultural census or geographical mapping. It is acknowledged that in several instances this information can be obtained at group level only, and an important challenge will be to obtain data at individual level and/or on individual habits.

Biobanks also constitute new data sources from healthy or diseased populations. They consist of an organised collection of human biological specimens and associated information stored for diverse research purposes. These biosamples are available for application of novel technologies with potential for generating data valuable for exposure assessment or exposure reconstruction. If studies' design and conduct are harmonised, data and samples can be shared between biobanks to promote powerful pooled analyses and replications studies (Burton et al., 2010).

Large scale epidemiological studies with deep phenotyping provide also unprecedented opportunities to link well phenotyped study participants with the aforementioned data. For example, UK Biobank, has recruited over 500,000 individuals with questionnaire, medical history and physical measurements data as well as stored blood and urine samples with available genome wide association data for all 500,000 participants, and linkage to Hospital Episode Statistics, national registry data and primary care records. To gain information on air pollution and noise levels, the postcode of participants has been linked to air pollution or noise estimates. In addition, piloting of personal exposure monitoring will take place in order to collect individual level data on these exposures. These approaches could be extended to gain information on pesticide exposure, either through geographical linkage, linkage with purchasing and occupational registries, and personal exposure monitoring. Similar biobanks exist in many other EU countries (http://www.bbmri-eric.eu/BBMRI-ERIC has collected most EU studies).

8. Overall recommendations

8.1. Recommendations for single epidemiological studies:

The following recommendations for improving epidemiological studies are aimed to conform to the 'recognised standards' mentioned in Regulation (EU) No 1107/2009 to make them of particular value to risk assessment of pesticides ('where available, and supported with data on levels and duration of exposure, and conducted in accordance with recognised standards, epidemiological studies are of particular value and must be submitted'). Accordingly, these recommendations can indeed not be considered as a practical guidance for researchers on how to conduct such studies, but for those who are planning to conduct a study for further use in pesticide risk assessment.

a) Study design (including confounding)

1) Since prospective epidemiological designs provide stronger evidence for causal inference, these studies are encouraged over the other designs for pesticide risk assessment.
2) Future epidemiological studies should be conducted using the appropriate sample size in order to properly answer the question under investigation. A power analysis should thus be performed at the study design stage.
3) Future studies should take into consideration heterogeneity, subpopulations, exposure windows and susceptibility periods and conditions (pregnancy, development, diseases, etc.).

[18] National Academies of Sciences, Engineering, and Medicine; Division on Earth and Life Studies; Board on Environmental Studies and Toxicology; Committee on Incorporating 21st Century Science into Risk-Based Evaluations. Washington (DC): National Academies Press (US); 2017 Jan.

4) A wide range of potential confounding variables (including co-exposure to other chemicals, lifestyle, socioeconomic factors, etc.) should be measured or accounted for during the design stage (e.g. matching) of the study.
5) Consideration of host factors that may influence toxicity and act as effect modifiers. These will include genetic polymorphisms data (e.g. paraoxonase-1 genotype) or nutritional factors (e.g. iodine status) among others.
6) Collaboration between researchers is encouraged to build-up consortia that enhance the effectiveness of individual cohorts.

Collection and appropriately storage of relevant biological material should be undertaken for future exposure assessment, including the use of novel technologies.

b) Exposure (measurement, data transformation for reporting and statistical analysis):

1) Collection of specific information on exposure should avoid as far as possible broad definitions of exposure, non-specific pesticide descriptions and broad exposures classifications such as 'never' *vs.* 'ever' categories. Nevertheless, these categories may be valuable under certain circumstances, e.g. to anticipate a class effect.
2) Studies which only look at broad classes of pesticides (generic groups of unrelated substances), or 'insecticides', 'herbicides', etc. or even just 'pesticides' in general are of much less use (if any) for risk assessment. Studies that investigate specific named pesticides and co-formulants are more useful for risk assessment.
3) Pesticides belonging to the same chemical class or eliciting the same mode of toxic action or toxicological effects might be grouped in the same category. Further refinement with information on frequency, duration and intensity of exposure might help in estimating exposure patterns.
4) In occupational epidemiology studies, operator and worker behaviour and proper use of PPE should be adequately reported as these exposure modifiers may significantly change exposures and thereby potential associations.
5) Improving the accuracy of exposure measurement is increasingly important, particularly for cohort studies. Long-term cohort studies which cover the etiologically relevant time period should improve the accuracy of measures of exposures by use of repeated biologic measures or repeated updates of self-reported exposures.
6) Indirect measures of environmental exposure for wider populations, including records on pesticide use, registry data, GIS, geographical mapping, etc., as well as data derived from large databases (including administrative databases) may be valuable for exploratory studies. If these data are not available, records/registries should be initiated. Likewise, estimation of dietary exposure to pesticide from food consumption databases and levels of pesticide residues from monitoring programmes can be used as well. As with direct exposure assessment, each method of indirect measurement should be reviewed for risk of bias and misclassification and weighted appropriately.
7) Whenever possible, exposure assessment should use direct measurements of exposure to named pesticides in order to establish different levels of exposure (e.g. personal exposure metering/biological monitoring), possibly in conjunction with other methods of exposure assessment which are more practicable or even necessary for large studies and historical exposures. New studies should explore novel ways of personal exposure monitoring. Results should be expressed using standardised units to normalise exposure across populations
8) The characterisation of exposure assessment over time can benefit by undertaken a more comprehensive exposure monitoring strategy coupled with information on exposure determinants over a longer time period collected from questionnaires or job-exposure matrices supported by biomonitoring data. Exposure assessment models can be comprehensively supported by HBM studies, which would allow identification of the critical exposure parameters. If such case, adjustments can then be made to the parameter assumptions within the models, leading to more realistic evaluations of exposure.
9) The use of the exposome concept and metabolomics in particular hold great promise for next-generation epidemiological studies both for better exposure measurement (biomarkers of exposure), for identification of vulnerable subpopulations and for biological interpretation of toxicity pathways (biomarkers of disease).

10) Improved knowledge on exposure (and toxicity) to pesticide mixtures will be beneficial for comprehensive risk assessment. Consideration of the joint action of combined exposures to multiple pesticides acting on common targets, or eliciting similar adverse effects, is relevant for cumulative risk assessment. This requires all the components of the mixture to be known as well as an understanding of the MoA, dose–response characteristics and potential interactions between components. Characterisation of the exposure is a key element for combined exposure to multiple pesticides where the pattern and magnitude of exposure changes over time.

c) Adverse Outcomes (measurement, data transformation for reporting and statistical analysis):

1) Self-reported health outcomes should be avoided or confirmed by independent, blinded assessment of disease status by a medical expert assigned to the study.
2) Outcomes under study should be well defined and surrogate endpoints should be avoided unless they have been validated. Care must be taken when definitions of diseases and subclasses of diseases change over time (cancer, neurodegenerative disorders, etc.).
3) Use should be made of biological markers of early biological effect to improve the understanding of the pathogenesis of diseases. These quantitative biological parameters from mechanistic toxicology will enhance the usefulness of epidemiology because they improve the study sensitivity, reduce misclassification and enhance human relevance as compared to findings from studies in experimental animals. Since these refined endpoints are early events in the toxicodynamic pathway and often measured on a continuous scale, they might be preferable to more overt and traditional outcomes.
4) The use of biomarkers of effect may be helpful in assessing aggregate exposure to pesticides and informing cumulative risk assessment.
5) Developing read across methods allowing health outcomes to be identified using epidemiological studies and to link acute and chronic incidents records with experimental findings.

d) Statistical (descriptive statistics, modelling of exposure–effect relationship):

1) Statistical analysis should be based on *a priori* defined analytical (statistical) protocols, to avoid post hoc analyses for exploratory studies and report all the results, regardless of whether they are statistically significant or not.
2) Data should be reported in such a way that permit, where appropriate, mathematical modelling to estimate individual/population exposures and dose–response assessment irrespective of whether direct or indirect measures are used.
3) Reports should include both unadjusted and adjusted proportions and rates of outcome of interest across studies that are based on underlying populations with different structure of relevant factors and exposures.
4) Possible relevant factors, and their role in the exposure–health outcome relationship, should be carefully identified, accurately measured and thoroughly assessed. Most often, relevant factors have been screened as potential confounders. When confounding effects were detected, these needed to be adjusted for using appropriate statistical methods that include sensitivity analysis.
5) Potentially useful analytical approaches, such as propensity score matching, mediation analyses, and causal inference are encouraged to be applied in pesticide epidemiology.
6) When the association between a given pesticide exposure and a disease is found to be statistically significant, particularly in (presumed) low powered studies, it would be general good practice to perform a power analysis/design calculation to determine the degree to which the statistically significant effect size estimate (e.g. OR or RR) may be artificially inflated or magnified.[19]

[19] Additional information on power and sample size recommendations and related issues including effect size magnification and design calculations are provided in Annex D to this report. Specifically, a power calculation requires 3 values to be clearly reported by epidemiological studies: (i) the number of subjects in the non-exposed group (including individuals with and without the disease of interest); (ii) the number of subjects in the exposed group (also including individuals with and without the disease of interest); (iii) the number of diseased subjects in the non-exposed group.

e) Reporting of results:

1) These should follow practices of good reporting of epidemiological research outlined in the STROBE statement and in the EFSA guideline on statistical reporting (EFSA, 2014b) and include the further suggestions identified in this Opinion including effect size inflation estimates.
2) Although some epidemiological research will remain exploratory and post hoc in nature, this should be acknowledged and supported by appropriate statistical analysis.
3) Epidemiological studies are encouraged to provide access to raw data for further investigations and to deposit their full results and scripts or software packages used for analyses.
4) Report, or deposit using online sources, all results along with scripts and statistical tools used to allow the reproducibility of results to be tested.
5) Report all sources of funding and adequately report financial and other potential conflicts of interest.

As a general recommendation, the PPR Panel encourages development of guidance for epidemiological research in order to increase its value, transparency and accountability for risk assessment.[20] An increased quality of epidemiological studies, together with responsible research conduct and scientific integrity, will benefit the incorporation of these studies into risk assessment.

8.2. Surveillance

1) Increase the reporting of acute and chronic incidents by setting up post-marketing surveillance programmes (occupational and general population) as required by article 7 of EU directive 2009/128; this should be fulfilled by developing surveillance networks with occupational health physicians and by boosting the collaboration between national authorities dealing with PPP and poison control information centres.
2) Develop a valid method for assessing the weight/strength of the causal relationship ('imputability') for acute and chronic incidents, and develop glossaries and a thesaurus to support harmonised reporting between EU member states.
3) Harmonised data from member states should be gathered at the EU level and examined periodically by the Commission/EFSA and a report should be released focussing on the most relevant findings.
4) Develop an EU-wide vigilance framework for pesticides.
5) There is scope for training improvements regarding pesticide toxidromes in toxicology courses for medical and paramedical staff responsible for diagnostic decisions, data entry and management.

8.3. Meta-analysis of multiple epidemiological studies

1) Evidence from epidemiological studies might be pooled by taking into account a thorough evaluation of the methods and biases of individual studies, an assessment of the degree of heterogeneity among studies, development of explanations underlying any heterogeneity and a quantitative summary of the evidence (provided that it is consistent).
2) For every evidence synthesis effort, studies should be reviewed using relevant risk of bias tools. Studies with different designs, or with different design features, may require (some) different questions for risk of bias assessments.
3) Evidence syntheses should not be restricted to specific time frames; they should include the totality of evidence. These efforts are more relevant if focused on specific health outcome or disease categories.
4) In evidence synthesis efforts, beyond the quantitative synthesis of the effect sizes, there should be consideration on the calculated predictive intervals, small study effects and asymmetry bias, conflicts of interest, confounding, excess significance bias,[21] and heterogeneity estimates.

[20] An example is the guideline developed by the Dutch Society for Epidemiology on responsible epidemiologic Research Practice (2017).

[21] Excess significance bias refers to the situation in which there are too many studies with statistically significant results in the published literature on a particular outcome. This pattern suggests strong biases in the literature, with publication bias, selective outcome reporting, selective analyses reporting, or fabricated data being possible explanations (Ioannidis and Trikalinos, 2007).

5) In the presence of heterogeneity, studies with highly selected populations, albeit unrepresentative of their respective populations, may prove valuable and deserve consideration as they may represent genuine and not statistical heterogeneity.
6) A more consistent reporting such as for age, race and gender across studies would enhance the meta-analyses.
7) Where quantitative data of individual pesticides are available from epidemiological studies, they can be combined or pooled for dose–response modelling, which could enable development of quantitative risk estimates and points of departure (BMDL, NOAEL).
8) International consortium of cohort studies should be encouraged to support data pooling to study disease–exposure associations that individual cohorts do not have sufficient statistical power to study (e.g. AGRICOH).

8.4. Integration of epidemiological evidence with other sources of information

1) All lines of evidence (epidemiology, animal, *in vitro* data) should be equally scrutinised for biases.
2) Validated and harmonised methods should be developed to combine observational studies, animal/basic science studies and other sources of evidence for risk assessment.
3) Experimental and human data should both contribute to hazard identification and to dose–response assessment.
4) A systematic integration of data from multiple lines of evidence should be based on a WoE analysis accounting for relevance, consistency and biological plausibility using modified Bradford Hill criteria. The principles underlying this framework are described in Section 7.2 and summarised in Figure 5.
5) Epidemiological findings should be integrated with other sources of information (data from experimental toxicology, mechanism of action/AOP) by using a WoE approach. An integrated and harmonised approach should be developed by bringing together animal, mechanistic and human data in an overall WoE framework in a systematic and consistent manner.
6) The AOP framework offers a structured platform for the integration of various kinds of research results.
7) Animal, *in vitro* data and human data should be assessed as a whole for each endpoint. A conclusion can be drawn as to whether the results from the experiments are confirmed by human data for each endpoint and this could be included in the RARs.

9. Conclusions

This Scientific Opinion is intended to help the peer review process during the renewal of pesticides authorisation (and, where possible, during the approval process) under Regulation 1107/2009 which requires a search of the scientific peer-reviewed open literature, including existing epidemiological studies. These are more suitable for the renewal process of active substances, also in compliance with Regulation 1141/2010, which indicates that the dossiers submitted for renewal should include new data relevant to the active substance.

The four key elements of the terms of reference are repeated below and the parts of the text addressing the individual terms are identified in order. As they follow from the text passages grouped with each of the ToRs the recommendations relevant to each of the ToRs are also indicated as follows.

'The PPR Panel will discuss the associations between pesticide exposure and human health effects observed in the External scientific report (Ntzani et al., 2013) and how these findings could be interpreted in a regulatory pesticide risk assessment context. Hence, the PPR Panel will systematically assess the epidemiological studies collected in the report by addressing major data gaps and limitations of the studies and provide recommendations thereof'.

'The PPR Panel will specifically':

1) Collect and review all sources of gaps and limitations, based on (but not necessarily limited to) those identified in the External Scientific report in regard to the quality and relevance of the available epidemiological studies. Responses in Section 3 pp. 20–24, Section 5.2 pp. 33–35: no Recommendations appropriate.
2) Based on the gaps and limitations identified in point 1, propose potential refinements for future epidemiological studies to increase the quality, relevance and reliability of the findings

and how they may impact pesticide risk assessment. This may include study design, exposure assessment, data quality and access, diagnostic classification of health outcomes, and statistical analysis. Responses in Section 4 pp 24–33: recommendations in Sections 8.1, 8.2 and 8.3 pp. 54–58.

3) Identify areas in which information and/or criteria are insufficient or lacking and propose recommendations for how to conduct pesticide epidemiological studies in order to improve and optimise the application in risk assessment. These recommendations should include harmonisation of exposure assessment (including use of biomonitoring data), vulnerable population sub-groups and/or health outcomes of interest (at biochemical, functional, morphological and clinical level) based on the gaps and limitations identified in point 1. Responses in Sections 4.2–4.5 pp. 27–33, Section 5.3 pp. 36: recommendations in Section 8.1 c) 1–4, pp. 56.
4) Discuss how to make appropriate use of epidemiological findings in risk assessment of pesticides during the peer review process of draft assessment reports, e.g. WoE as well as integrating the epidemiological information with data from experimental toxicology, AOPs, mechanism of actions, etc. Responses in Sections 6.2 and 6.3 pp. 37–45 and 7 pp. 45–54: Responses in Section 8.4 pp. 58.

As explained above, appropriate epidemiological data and post-approval surveillance may usefully contribute to the risk assessment framework by hazard identification, and – with methodological improvements – hazard characterisation. It can be improved by contributions from WoE analysis, Uncertainty analysis, and identification and estimation of biases. It is the responsibility of applicants to collect the available relevant literature, to consider its relevance and quality using relevant EFSA criteria including those for systematic review and to introduce discussion of the outcomes within the DAR, RAR and post-approval frameworks that are prescribed under EU law.

The definition of appropriate quality will require analysis of sample size, statistical procedures, estimates of effect size inflation, assessment of biases and their contribution to the conclusions drawn.

The nature of the studies will require consideration at all relevant points in the risk assessment process so that for example epidemiological data on reproductive topics will be considered alongside laboratory animal studies designed to reveal reproductive effects and in the context of recommendation for labelling for reproductive toxicity (for ECHA).

Unless there is history of use in countries outside the EU, the relevant epidemiological studies will be restricted in their effect on the DAR but the RAR and Surveillance framework is potentially able to benefit from epidemiology progressively as time after first approval passes and from prior use of Active Ingredients in other jurisdictions. It is recommended that RAR and surveillance protocols should reflect this difference.

The specific recommendations listed above follow from detailed arguments based on an analysis of present and foreseen **s**trengths **w**eaknesses **o**pportunities and **t**hreats related to the use of epidemiological data in risk assessment. Broadly these are as follows:

Strengths. Include:

- The fact that the evidence concerns human specific risks.
- That health outcomes are integrated measures of the effects of all exposure to toxins.
- The ability to elicit subjective experience from potentially affected people.

Weaknesses. Include:

- The exposures to pesticides are usually complex; contribution of a specific active ingredient is not easily deciphered.
- The exposures occur in various settings where precisely controlled conditions are lacking.
- Most data reflect the responses of mixed populations.
- Many data show low level associations that are inconsistently repeatable and require sophisticated analysis.

Opportunities. Despite the range of limitations described in this Opinion, which apply to many available published epidemiological studies, there are opportunities to benefit risk assessment of pesticides. These include:

- The access to very large numbers of potentially exposed individuals for studies that may reveal subtle health effects and reveal the experience of sensitive sub-groups.

- The prospect of improving exposure estimation using biomonitoring and new molecular approaches to establish tissue burdens of potential toxins and their residues.
- The possibility of fully integrating human data into the conventional risk assessment based on responses in laboratory animals.
- Utilising WoE, AOP, Expert judgement, Expert Knowledge Elicitation (EKE) and Uncertainty Analysis to evaluate differences in the quality of potentially relevant data.
- The opportunity to engage professional epidemiologists and statisticians to refine interpretation of epidemiological findings and to recommend improved designs to tackle difficult areas such as chronic and combined exposure risks and dose–response data.
- A major information technology opportunity exists in pooling data from a variety of national sources. Once the relevant legal, methodological and ethical issues are overcome much more valuable data can be collected. When this data is made available, in a form that can be used in a 'big data' setting for societal benefit there will be potential for significant improvements in epidemiological studies. First, however, it will be necessary to preserve individual privacy and essential commercial confidentiality. Once these obstacles are overcome the statistical power of epidemiological studies can be improved and applied to identify and possibly characterise hazards better. These aims can be realised effectively by agreed actions at a high EU level. Interstate approval for providing data and interactive platforms will need to be backed by harmonisation of population health information, food consumption data, active substance and co-formulant spatial and temporal application data. Such rich data can be expected to assist in increasing consistency, a criterion that strengthens evidence of causality and reliability. It promises larger sample sizes for epidemiological studies that will be better able to identify vulnerable groups that may require special protection from pesticide toxicity.

Threats. Include:

- Widespread perception of risk levels to the human population or to wildlife and the environment that are unrealistic and that cause negative consequences in societies.
- Poor experimental design yielding false positive or false negative conclusions that undermine data from other valid sources.
- Failure to respond to emerging risks as a result of ineffective surveillance or unwillingness to make appropriate anonymised data available for societal benefit.
- Waste of data through failure to collect appropriate information regarding exposure (specifically occupational exposure) by registries (cancer or congenital anomalies) or surveillance programmes which hinders linking health outcomes to exposure.
- Waste of data through failure to harmonise diagnostic criteria, failure to record data in a sufficiently detailed combinable form for integrated analysis, poor training of medical and paramedical staff in relevant toxidromes that will allow optimum quality of data entered into Health Statistics Databases.

References

Adami HO, Berry SC, Breckenridge CB, Smith LL, Swenberg JA, Trichopoulos D, Weiss NS and Pastoor TP, 2011. Toxicology and epidemiology: improving the science with a framework for combining toxicological and epidemiological evidence to establish causal inference. Toxicology Sciences, 122, 223–234.

Amler RW, Barone Jr S, Belger A, Berlin Jr CM, Cox C, Frank H, Goodman M, Harry J, Hooper SR, Ladda R, LaKind JS, Lipkin PH, Lipsitt LP, Lorber MN, Myers G, Mason AM, Needham LL, Sonawane B, Wachs TD and Yager JW, 2006. Hershey Medical Center Technical Workshop Report: optimizing the design and interpretation of epidemiologic studies for assessing neurodevelopmental effects from in utero chemical exposure. Neurotoxicology, 27, 861–874.

Bengtson AM, Westreich D, Musonda P, Pettifor A, Chibwesha C, Chi BH, Vwalika B, Pence BW, Stringer JS and Miller WC, 2016. Multiple overimputation to address missing data and measurement error: application to HIV treatment during pregnancy and pregnancy outcomes. Epidemiology, 27, 642–650.

Bevan R, Brown T, Matthies F, Sams C, Jones K, Hanlon J and La Vedrine M, 2017. Human Biomonitoring data collection from occupational exposure to pesticides. EFSA supporting publication 2017:EN-1185, 207 pp.

Bottai M, 2014. Lessons in biostatistics: inferences and conjectures about average and conditional treatment effects in randomized trials and observational studies. Journal of Internal Medicine, 276, 229–237.

Budtz-Jørgensen E, Keiding N and Grandjean P, 2001. Benchmark dose calculation from epidemiological data. Biometrics, 57, 698–706.

Budtz-Jørgensen E, Keiding N and Grandjean P, 2004. Effects of exposure imprecision on estimation of the benchmark dose. Risk Analysis, 24, 1689–1696.

Buonsante VA, Muilerman H, Santos T, Robinson C and Tweedale AC, 2014. Risk assessment's insensitive toxicity testing may cause it to fail. Environmental Research, 135, 139–147.
Burton PR, Fortier I and Knoppers BM, 2010. The global emergence of epidemiological biobanks: opportunities and challenges. In: Khoury M, Bedrosian S, Gwinn M, Higgins J, Ioannidis J and Little J (eds.). Human Genome Epidemiology. Building the evidence for using genetic information to improve health and prevent disease. 2nd Edition, Oxford University Press, Oxford. pp. 77–99.
Choi J, Polcher A and Joas A, 2016. Systematic literature review on Parkinson's disease and Childhood Leukaemia and mode of actions for pesticides. EFSA supporting publication 2016:EN-955, 256 pp. Available online: http://www.onlinelibrary.wiley.com/doi/10.2903/sp.efsa.2016.EN-955/pdf
Coble J, Thomas KW, Hines CJ, Hoppin JA, Dosemeci M, Curwin B, Lubin JH, Beane Freeman LE, Blair A, Sandler DP and Alavanja MC, 2011. An updated algorithm for estimation of pesticide exposure intensity in the agricultural health study. International Journal of Environmental Research and Public Health, 8, 4608–4622.
Coggon D, 1995. Questionnaire based exposure assessment methods. Science of the Total Environment, 168, 175–178.
Cornelis C, Schoeters G, Kellen E, Buntinx F and Zeegers M, 2009. Development of a GIS-based indicator for environmental pesticide exposure and its application to a Belgian case-control study on bladder cancer. International Journal of Hygiene and Environmental Health, 212, 172–185.
la Cour JL, Brok J and Gøtzsche PC, 2010. Inconsistent reporting of surrogate outcomes in randomised clinical trials: cohort study. BMJ, 341, c3653.
DeBord DG, Burgoon L, Edwards SW, Haber LT, Kanitz MH, Kuempel E, Thomas RS and Yucesoy B, 2015. Systems biology and biomarkers of early effects for occupational exposure limit setting. The Journal of Occupational and Environmental Hygiene, 12(Suppl 1), S41–S54.
Dionisio KL, Chang HH and Baxter LK, 2016. A simulation study to quantify the impacts of exposure measurement error on air pollution health risk estimates in copollutant time-series models. Environmental Health, 15, 114.
DSE (Dutch Society for Epidemiology), 2017. Responsible Epidemiologic Research Practice (RERP). A guideline developed by the RERP working group of the Dutch Society for Epidemiology, 2017 (available at https://www.epidemiologie.nl/home.html, https://epidemiologie.nl/fileadmin/Media/docs/Onderzoek/Responsible_Epidemiologic_Research_Practice.2017.pdf)
ECETOC (European Centre for Ecotoxicology and Toxicology of Chemicals), 2009. Framework for the Integration of Human and Animal Data in Chemical Risk Assessment. Technical Report No. 104. Brussels. Available online: http://www.ecetoc.org/uploads/Publications/documents/TR%20104.pdf
ECHA/EFSA, 2014. Workshop on Mode of action and Human relevance framework in the context of classification and labelling (CLH) and regulatory assessment of biocides and pesticides. November 2014. Available online: https://echa.europa.eu/documents/10162/22816050/moaws_workshop_proceedings_en.pdf/a656803e-4d97-438f-87ff-fc984cfe4836
EFSA (European Food Safety Authority), 2004. Opinion of the Scientific Panel on Dietetic Products, Nutrition and Allergies on a request from the Commission related to the presence of trans fatty acids in foods and the effect on human health of the consumption of trans fatty acids. EFSA Journal 2004;81, 1–49 pp. https://doi.org/10.2903/j.efsa.2004.81
EFSA (European Food Safety Authority), 2009a. Scientific Opinion of the Panel on Contaminants in the Food Chain on a request from the European Commission on cadmium in food. EFSA Journal 2009;980, 1–139 pp. https://doi.org/10.2903/j.efsa.2009.980
EFSA (European Food Safety Authority Panel on Contaminants in the Food Chain CONTAM), 2009b. Scientific Opinion on arsenic in food. EFSA Journal 2009;7(10):1351, 199 pp. https://doi.org/10.2903/j.efsa.2009.1351
EFSA (European Food Safety Authority), 2010a. Application of systematic review methodology to food and feed safety assessments to support decision making. EFSA Journal 2010;8(6):1637, 90 pp. https://doi.org/10.2903/j.efsa.2010.1637
EFSA (European Food Safety Authority) Panel on Contaminants in the Food Chain (CONTAM), 2010b. Scientific Opinion on Lead in Food. EFSA Journal 2010;8(4):1570, 151 pp. https://doi.org/10.2903/j.efsa.2010.1570
EFSA (European Food Safety Authority), 2011a. Submission of scientific-peer reviewed open literature for the approval of pesticide active substances under Regulation (EC) No 1107/2009. EFSA Journal 2011;9(2):2092, 49 pp. https://doi.org/10.2903/j.efsa.2011.2092
EFSA (European Food Safety Authority), 2011b. Statistical significance and biological relevance. EFSA Journal 2011;9(9):2372, 17 pp. https://doi.org/10.2903/j.efsa.2011.2372
EFSA (European Food Safety Authority), 2012a. Scientific Opinion on risk assessment terminology. EFSA Journal 2012;10(5):2664, 43 pp. https://www.efsa.europa.eu/en/efsajournal/pub/2664
EFSA (European Food Safety Authority Panel on Contaminants in the Food Chain CONTAM), 2012b. Scientific Opinion on the risk for public health related to the presence of mercury and methylmercury in food. EFSA Journal 2012;10(12):2985, 241 pp. https://doi.org/10.2903/j.efsa.2012.2985
EFSA (European Food Safety Authority), 2013a. Scientific Opinion on the identification of pesticides to be included in cumulative assessment groups on the basis of their toxicological profile. EFSA Journal 2013;11(7):3293, 131 pp. https://doi.org/10.2903/j.efsa.2013.3293

EFSA (European Food Safety Authority), 2013b. Scientific Opinion on the relevance of dissimilar mode of action and its appropriate application for cumulative risk assessment of pesticides residues in food. EFSA Journal 2013;11(12):3472, 40 pp. https://doi.org/10.2903/j.efsa.2013.3472
EFSA (European Food Safety Authority), 2014a. Conclusion on the peer review of the pesticide human health risk assessment of the active substance chlorpyrifos. EFSA Journal 2014;12(4):3640, 34 pp. https://doi.org/10.2903/j.efsa.2014.3640
EFSA (European Food Safety Authority), 2014b. Guidance on statistical reporting. EFSA Journal 2014;12(12):3908, 18 pp. https://doi.org/10.2903/j.efsa.2014.3908
EFSA (European Food Safety Authority), 2015a. Stakeholder Workshop on the use of epidemiological data in pesticide risk assessment. EFSA supporting publication 2015:EN-798, 8 pp. Available online: https://www.efsa.europa.eu/en/supporting/pub/798e
EFSA (European Food Safety Authority), 2015b. Increasing robustness, transparency and openness of scientific assessments – Report of the Workshop held on 29–30 June 2015 in Brussels. EFSA supporting publication 2015:EN-913. 29 pp. Available online: http://www.efsa.europa.eu/sites/default/files/corporate_publications/files/913e.pdf
EFSA (European Food Safety Authority), 2015c. Conclusion on the peer review of the pesticide risk assessment of the active substance glyphosate. EFSA Journal 2015;13(11):4302, 107 pp. https://doi.org/10.2903/j.efsa.2015.4302
EFSA PPR Panel (European Food Safety Authority Panel on Plant Protection Products and their Residues), 2017. Scientific Opinion on the investigation into experimental toxicological properties of plant protection products having a potential link to Parkinson's disease and childhood leukaemia. EFSA Journal 2017;15(3):4691, 325 pp. https://doi.org/10.2903/j.efsa.2017.4691
EFSA Scientific Committee (European Food Safety Authority Scientific Committee), 2017a. Guidance on the assessment of the biological relevance of data in scientific assessments. EFSA Journal 2017;15(8):4970, 73 pp. https://doi.org/10.2903/j.efsa.2017.4970
EFSA Scientific Committee (European Food Safety Authority Scientific Committee), 2017b. Guidance on the use of the weight of evidence approach in scientific assessments. EFSA Journal 2017;15(8):4971, 69 pp. https://doi.org/10.2903/j.efsa.2017.4971.
EFSA Scientific Committee (European Food Safety Authority Scientific Committee), 2017c. Update: guidance on the use of the benchmark dose approach in risk assessment. EFSA Journal 2017;15(1): 4658, 41 pp. https://doi.org/10.2903/j.efsa.2017.4658
von Elm E, Altman DG, Egger M, Pocock SJ and Gøtzsche PC, Vandenbroucke JP and STROBE Initiative, 2007. Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. BMJ, 335, 806–808.
Esch EW, Bahinski A and Huh D, 2015. Organs-on-chips at the frontiers of drug discovery. Nature Reviews. Drug Discovery, 14, 248–260.
Fedak KM, Bernal A, Capshaw ZA and Gross S, 2015. Applying the Bradford Hill criteria in the 21st century: how data integration has changed causal inference in molecular epidemiology. Emerging Themes in Epidemiology, 30, 14.
Gibson SB, Downie JM, Tsetsou S, Feusier JE, Figueroa KP, Bromberg MB, Jorde LB and Pulst SM, 2017. The evolving genetic risk for sporadic ALS. Neurology, 89, 226–233.
Gómez-Martín A, Hernández AF, Martínez-González LJ, González-Alzaga B, Rodríguez-Barranco M, López-Flores I, Aguilar-Garduno C and Lacasana M, 2015. Polymorphisms of pesticide-metabolizing genes in children living in intensive farming communities. Chemosphere, 139, 534–540.
González-Alzaga B, Hernández AF, Rodríguez-Barranco M, Gómez I, Aguilar-Garduño C, López-Flores I, Parrón T and Lacasaña M, 2015. Pre- and postnatal exposures to pesticides and neurodevelopmental effects in children living in agricultural communities from South-Eastern Spain. Environment International, 85, 229–237.
Greenland S and Longnecker MP, 1992. Methods for trend estimation from summarized dose-response data, with applications to meta-analysis. American Journal of Epidemiology, 135, 1301–1309.
Greenland S and O'Rourke K, 2008. Meta-analysis. In: Rothman K, Greenland S and Lash T (eds). *Modern Epidemiology. 3*. Lippincott Williams & and Wilkins, Philadelphia. pp. 652–682.
Greenland S, Senn SJ, Rothman KJ, Carlin JB, Poole C, Goodman SN and Altman DG, 2016. Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. European Journal of Epidemiology, 31, 337–350.
Grimes DA and Schulz KF, 2005. Surrogate end points in clinical research: hazardous to your health. Obstetrics and Gynecology, 105, 1114–1118.
Gustafson P and McCandless LC, 2010. Probabilistic approaches to better quantifying the results of epidemiologic studies. International Journal of Environmental Research and Public Health, 7, 1520–1539.
Hernández AF, González-Alzaga B, López-Flores I and Lacasaña M, 2016. Systematic reviews on neurodevelopmental and neurodegenerative disorders linked to pesticide exposure: methodological features and impact on risk assessment. Environment International, 92–93, 657–679.
Higgins JP, 2008. Commentary: heterogeneity in meta-analysis should be expected and appropriately quantified. International Journal of Epidemiology, 37, 1158–1160.

Hill AB, 1965. The environment and disease: association or causation? Proceedings of the Royal Society of Medicine, 58, 295–300.
Hines CJ, Deddens JA, Coble J, Kamel F and Alavanja MC, 2011. Determinants of captan air and dermal exposures among orchard pesticide applicators in the Agricultural Health Study. Annals of Occupational Hygiene, 55, 620–633.
Hoffmann S, de Vries RBM, Stephens ML, Beck NB, Dirven HAAM, Fowle JR 3rd, Goodman JE, Hartung T, Kimber I, Lalu MM, Thayer K, Whaley P, Wikoff D and Tsaioun K, 2017. A primer on systematic reviews in toxicology. Archives of Toxicology, 91, 2551–2575.
Höfler M, 2005. The Bradford Hill considerations on causality: a counterfactual perspective. Emerging Themes in Epidemiology, 2, 11.
IEA (International Epidemiological Association), 2007. Good Epidemiological Practice (GEP) 2007. Available online: http://ieaweb.org/good-epidemiological-practice-gep/
Imbens G and Rubin D, 2015. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press, New York, NY.
INSERM, 2013. Pesticides. Effets sur la santé. Collection expertise collective, Inserm, Paris, 2013.
Ioannidis JP and Trikalinos TA, 2007. An exploratory test for an excess of significant findings. Clinical Trials, 4, 245–253.
Jurek AM, Greenland S, Maldonado G and Church TR, 2005. Proper interpretation of non-differential misclassification effects: expectations vs observations. International Journal of Epidemiology, 34, 680–687.
Kaltenhäuser J, Kneuer C, Marx-Stoelting P, Niemann L, Schubert J, Stein B and Solecki R, 2017. Relevance and reliability of experimental data in human health risk assessment of pesticides. Regulatory Toxicology and Pharmacology, 88, 227–237.
Karabatsos G, Talbott E and Walker SG, 2015. A Bayesian nonparametric meta-analysis model. Research Synthesis Methods, 6, 28–44.
Kavvoura FK, Liberopoulos G and Ioannidis JP, 2007. Selection in reported epidemiological risks: an empirical assessment. PLoS Medicine, 4, e79.
Lachenmeier DW, Kanteres F and Rehm J, 2011. Epidemiology-based risk assessment using the benchmark dose/margin of exposure approach: the example of ethanol and liver cirrhosis. International Journal of Epidemiology, 40, 210–218.
LaKind JS, Sobus JR, Goodman M, Barr DB, Furst P, Albertini RJ, Arbuckle TE, Schoeters G, Tan YM, Teequarden J, Tornero-Velez R and Weisel CP, 2014. A proposal for assessing study quality: biomonitoring, environmental epidemiology, and short-lived chemicals (BEES-C) instrument. Environmental International, 73, 195–207.
LaKind JS, Goodman M, Barr DB, Weisel CP and Schoeters G, 2015. Lessons learned from the application of BEES-C: systematic assessment of study quality of epidemiologic research on BPA, neurodevelopment, and respiratory health. Environment International, 80, 41–71.
Landgren O, Kyle RA, Hoppin JA, Beane Freeman LE, Cerhan JR, Katzmann JA, Rajkumar SV and Alavanja MC, 2009. Pesticide exposure and risk of monoclonal gammopathy of undetermined significance in the Agricultural Health Study. Blood, 113, 6386–6391.
Larsson MO, Nielsen VS, Brandt CØ, Bjerre N, Laporte F and Cedergreen N, 2017. Quantifying dietary exposure to pesticide residues using spraying journal data. Food and Chemical Toxicology, 105, 407–428.
Lash TL, Fox MP and Fink AK, 2009. *Applying Quantitative Bias Analysis to Epidemiologic Data*. Springer, New York.
Lavelle KS, Robert Schnatter A, Travis KZ, Swaen GM, Pallapies D, Money C, Priem P and Vrijhof H, 2012. Framework for integrating human and animal data in chemical risk assessment. Regulatory Toxicology and Pharmacology, 62, 302–312.
London L, Coggon D, Moretto A, Westerholm P, Wilks MF and Colosio C, 2010. The ethics of human volunteer studies involving experimental exposure to pesticides: unanswered dilemmas. Environmental Health, 18, 50.
Maldonado G and Greenland S, 2002. Estimating causal effects. International Journal of Epidemiology, 31, 422–429.
Marx-Stoelting P, Braeuning A, Buhrke T, Lampen A, Niemann L, Oelgeschlaeger M, Rieke S, Schmidt F, Heise T, Pfeil R and Solecki R, 2015. Application of omics data in regulatory toxicology: report of an international BfR expert workshop. Archives of Toxicology, 89, 2177–2184.
McNamee R, 2003. Confounding and confounders. Occupational and Environmental Medicine, 60, 227–234.
Monson R, 1990. *Occupational Epidemiology*, 2nd Edition. CRC Press, Boca Ration, FL.
Muñoz-Quezada MT, Lucero BA, Barr DB, Steenland K, Levy K, Ryan PB, Iglesias V, Alvarado S, Concha C, Rojas E and Vega C, 2013. Neurodevelopmental effects in children associated with exposure to organophosphate pesticides: a systematic review. Neurotoxicology, 39, 158–168.
Nachman KE, Fox MA, Sheehan MC, Burke TA, Rodricks JV and Woodruff TJ, 2011. Leveraging epidemiology to improve risk assessment. Open Epidemiology Journal, 4, 3–29.
Nieuwenhuijsen MJ, 2015. Exposure assessment in environmental epidemiology. In: Vrijheid M (ed.). The Exposome-Concept and Implementation in Birth Cohorts Chapter 14. Oxford University Press.
NRC (National Research Council), 2007. Toxicity Testing in the 21st Century: A Vision and a Strategy. Washington, DC: The National Academies Press.
NRC (National Research Council), 2009. *Science and Decisions: Advancing Risk Assessment*. The National Academies Press, Washington, DC.

Ntzani EE, Chondrogiorgi M, Ntritsos G, Evangelou E and Tzoulaki I, 2013. Literature review on epidemiological studies linking exposure to pesticides and health effects. EFSA supporting publication 2013:EN-497, 159 pp.
OECD (Organisation for Economic Co-operation and Development), 2013. Guidance Document on Developing and Assessing Adverse Outcome Pathways. Series on Testing and Assessment, No. 184. Paris. Avilable online: http://search.oecd.org/officialdocuments/displaydocumentpdf/?cote=env/jm/mono%282013%296&doclanguage=en
Orford R, Crabbe H, Hague C, Schaper A and Duarte-Davidson R, 2014. EU alerting and reporting systems for potential chemical public health threats and hazards. Environment International, 72, 15–25.
Orford R, Hague C, Duarte-Davidson R, Settimi L, Davanzo F, Desel H, Pelclova D, Dragelyte G, Mathieu-Nolf M, Jackson G and Adams R, 2015. Detecting, alerting and monitoring emerging chemical health threats: ASHTIII. European Journal of Public Health, 25(supp 3), 218.
Orsini N, Li R, Wolk A, Khudyakov P and Spiegelman D, 2012. Meta-analysis for linear and nonlinear dose-response relations: examples, an evaluation of approximations, and software. American Journal of Epidemiology, 175, 66–73.
Oulhote Y and Bouchard MF, 2013. Urinary metabolites of organophosphate and pyrethroid pesticides and behavioral problems in Canadian children. Environmental Health Perspectives, 121, 1378–1384.
Pearce N, 2011. Registration of protocols for observational research is unnecessary and would do more harm than good. Occupational and Environmental Medicine, 68, 86–88.
Pearce N, 2012. Classification of epidemiological study designs. International Journal of Epidemiology, 41, 393–397.
Pearce N, Blair A, Vineis P, Ahrens W, Andersen A, Anto JM, Armstrong BK, Baccarelli AA, Beland FA, Berrington A, Bertazzi PA, Birnbaum LS, Brownson RC, Bucher JR, Cantor KP, Cardis E, Cherrie JW, Christiani DC, Cocco P, Coggon D, Comba P, Demers PA, Dement JM, Douwes J, Eisen EA, Engel LS, Fenske RA, Fleming LE, Fletcher T, Fontham E, Forastiere F, Frentzel-Beyme R, Fritschi L, Gerin M, Goldberg M, Grandjean P, Grimsrud TK, Gustavsson P, Haines A, Hartge P, Hansen J, Hauptmann M, Heederik D, Hemminki K, Hemon D, Hertz-Picciotto I, Hoppin JA, Huff J, Jarvholm B, Kang D, Karagas MR, Kjaerheim K, Kjuus H, Kogevinas M, Kriebel D, Kristensen P, Kromhout H, Laden F, Lebailly P, LeMasters G, Lubin JH, Lynch CF, Lynge E, 't Mannetje A, McMichael AJ, McLaughlin JR, Marrett L, Martuzzi M, Merchant JA, Merler E, Merletti F, Miller A, Mirer FE, Monson R, Nordby KC, Olshan AF, Parent ME, Perera FP, Perry MJ, Pesatori AC, Pirastu R, Porta M, Pukkala E, Rice C, Richardson DB, Ritter L, Ritz B, Ronckers CM, Rushton L, Rusiecki JA, Rusyn I, Samet JM, Sandler DP, de Sanjose S, Schernhammer E, Costantini AS, Seixas N, Shy C, Siemiatycki J, 2015. Silverman DT, Simonato L, Smith AH, Smith MT, Spinelli JJ, Spitz MR, Stallones L, Stayner LT, Steenland K, Stenzel M, Stewart BW, Stewart PA, Symanski E, Terracini B, Tolbert PE, Vainio H, Vena J, Vermeulen R, Victora CG, Ward EM, Weinberg CR, Weisenburger D, Wesseling C, Weiderpass E, Zahm SH. IARC monographs: 40 years of evaluating carcinogenic hazards to humans. Environmental Health Perspectives, 123, 507–514.
Raffaele KC, Vulimiri SV and Bateson TF, 2011. Benefits and barriers to using epidemiology data in environmental risk. The Journal of Epidemiology, 4, 99–105.
Raphael K, 1987. Recall bias: a proposal for assessment and control. International Journal of Epidemiology, 16, 167–170.
Rappaport SM, 2012. Biomarkers intersect with the exposome. Biomarkers, 17, 483–489.
Reich CG, Ryan PB and Schuemie MJ, 2013. Alternative outcome definitions and their effect on the performance of methods for observational outcome studies. Drug Safety, 36(Suppl 1), S181–S193.
Rothman KJ, 2002. *Epidemiology – An Introduction*. Oxford University Press, Oxford.
Rothman KJ and Greenland S, 1998. *Modern Epidemiology. 2*. Philadelphia: Lippincott Williams & Wilkins, 27 pp.
Rothman KJ, Greenland S and Lash TL, 2008. *Modern Epidemiology, 3rd Edition*. Lippincott Williams & Wilkins, Philadelphia, PA, USA.
Rushton L, 2011. Should protocols for observational research be registered? Occupational and Environmental Medicine, 68, 84–86.
Salerno J, Knoppers BM, Lee LM, Hlaing WW and Goodman KW, 2017. Ethics, big data and computing in epidemiology and public health. Annals of Epidemiology, 27, 297–301. https://doi.org/10.1016/j.annepidem.2017.05.002
Santacatterina M and Bottai M, 2015. Inferences and conjectures in clinical trials: a systematic review of generalizability of study findings. Journal of Internal Medicine, 279, 123–126. https://doi.org/10.1111/joim.12389
SCENIHR, 2012. Memorandum on the use of the scientific literature for human health risk assessment purposes – weighing of evidence and expression of uncertainty.
Simera I, Moher D, Hoey J, Schulz KF and Altman DG, 2010. A catalogue of reporting guidelines for health research. European Journal of Clinical Investigation, 40, 35–53.
Skelly AC, 2011. Probability, proof, and clinical significance. Evidence-Based Spine-Care Journal, 2, 9–11.
Spiegelman D, 2016. Evaluating Public Health Interventions: 4. the nurses' health study and methods for eliminating bias attributable to measurement error and misclassification. American Journal of Public Health, 106, 1563–1566.
Stang PE, Ryan PB, Dusetzina SB, Hartzema AG, Reich C, Overhage JM and Racoosin JA, 2012. Health outcomes of interest in observational data: issues in identifying definitions in the literature. Health Outcomes Research in Medicine, 3, e37–e44.
Thomas DC, 2009. *Statistical Methods in Environmental Epidemiology*. Oxford University Press, Oxford, UK.

Thomas KW, Dosemeci M, Coble JB, Hoppin JA, Sheldon LS, Chapa G, Croghan CW, Jones PA, Knott CE, Lynch CF, Sandler DP, Blair AE and Alavanja MC, 2010. Assessment of a pesticide exposure intensity algorithm in the agricultural health study. Journal of Exposure Science & Environmental Epidemiology, 20, 559–569.
Tsilidis KK, Panagiotou OA, Sena ES, Aretouli E, Evangelou E, Howells DW, Al-Shahi Salman R, Macleod MR and Ioannidis JP, 2013. Evaluation of excess significance bias in animal studies of neurological diseases. PLoS Biology, 11, e1001609.
Turner MC, Wigle DT and Krewski D, 2010. Residential pesticides and childhood leukemia: a systematic review and meta-analysis.
US EPA (United States Environmental Protection Agency), 2011. Chlorpyrifos: preliminary human health risk assessment for registration review, 30 June 2011, 159 pp.
US-EPA (U.S. Environmental Protection Agency), 2010a. Framework for incorporating human epidemiologic & incident data in health risk assessment (draft). Office of Pesticide Programs. Washington, DC, 2010.
US-EPA (U.S. Environmental Protection Agency), 2010b. Meeting Minutes of the FIFRA Scientific Advisory Panel Meeting on the Draft Framework and Case Studies on Atrazine, Human Incidents, and the Agricultural Health Study: Incorporation of Epidemiology and Human Incident Data into Human Health Risk Assessment. Arlington, Virginia, USA, April 22, 2010b. Available online: https://archive.epa.gov/scipoly/sap/meetings/web/pdf/020210minutes.pdf
US-EPA (U.S. Environmental Protection Agency), 2012. Guidance for considering and using open literature toxicity studies to support human health risk assessment. Office of Pesticide Programs. Washington, DC, 2012. Available online: http://www.epa.gov/pesticides/science/lit-studies.pdf
US-EPA (Environmental Protection Agency), 2016. Office of Pesticide Programs' Framework for Incorporating Human Epidemiologic & Incident Data in Risk Assessments for Pesticides December 28, 2016. Avilable online: https://www3.epa.gov/pesticides/EPA-HQ-OPP-2008-0316-DRAFT-0075.pdf
Vandenberg LN, Ågerstrand M, Beronius A, Beausoleil C, Bergman Å, Bero LA, Bornehag CG, Boyer CS, Cooper GS, Cotgreave I, Gee D, Grandjean P, Guyton KZ, Hass U, Heindel JJ, Jobling S, Kidd KA, Kortenkamp A, Macleod MR, Martin OV, Norinder U, Scheringer M, Thayer KA, Toppari J, Whaley P, Woodruff TJ and Rudén C, 2016. A proposed framework for the systematic review and integrated assessment (SYRINA) of endocrine disrupting chemicals. Environmental Health, 15, 74.
van den Brandt P, Voorrips L, Hertz-Picciotto I, Shuker D, Boeing H, Speijers G, Guittard C, Kleiner J, Knowles M, Wolk A and Goldbohm A, 2002. The contribution of epidemiology. Food and Chemical Toxicology, 40, 387–424.
Vinken M, 2013. The adverse outcome pathway concept: a pragmatic tool in toxicology. Toxicology, 312, 158–165.
Vlaanderen J, Moore LE, Smith MT, Lan Q, Zhang L, Skibola CF, Rothman N and Vermeulen R, 2010. Application of OMICS technologies in occupational and environmental health research: current status and projections. Occupational and Environmental Medicine, 67, 136–43.
WHO/IPCS (World Health Organization/International Programme on Chemical Safety), 2009. EHC 240: principles and methods for the risk assessment of chemicals in food.
Wilson SJ and Tanner-Smith EE, 2014. Meta-analysis in prevention science. In: Sloboda Z and Petras H (eds.). *Defining prevention science*. Advances in Prevention Science (vol. 1): Defining Prevention Science Springer, New York. pp. 431–452.
Youngstrom E, Kenworthy L, Lipkin PH, Goodman M, Squibb K, Mattison DR, Anthony LG, Makris SL, Bale AS, Raffaele KC and LaKind JS, 2011. A proposal to facilitate weight-of-evidence assessments: harmonization of Neurodevelopmental Environmental Epidemiology Studies (HONEES). Neurotoxicology and Teratology, 33, 354–359.
Zingone A and Kuehl WM, 2011. Pathogenesis of monoclonal gammopathy of undetermined significance and progression to multiple myeloma. Seminars in Hematology, 48, 4–12.

Glossary and Abbreviations

ADI	Acceptable daily intake. A measure of the amount of a pesticide in food or drinking water that can be ingested (orally) on a daily basis over a lifetime without an appreciable health risk.
ADME	Abbreviation used in pharmacology (and toxicology) for absorption, distribution, metabolism, and excretion of a chemical o pharmaceutical compound and describes its disposition within an organism.
AOP	Adverse Outcome Pathway. A structured representation of biological events leading to adverse effects relevant to risk assessment.
ARfD	Acute Reference Dose. An estimate of the amount a pesticide in food or drinking water (normally expressed on a body weight basis) that can be ingested in a period of 24 hours or less without appreciable health risks to the consumer on the basis of all known facts at the time of the evaluation.
Biomarker	Also known as 'biological marker'. A characteristic that is objectively measured and evaluated as an indication of normal biologic processes, pathogenic processes or pharmacologic responses to a therapeutic intervention

BMD	Benchmark Dose. A threshold dose or concentration that produces a predetermined change in response rate of an adverse effect (the benchmark response or BMR) compared to background. The lower 95% confidence limit is calculated (BMDL) to be further used as a point of departure to derive health-based reference values.
HBM	Human biomonitoring. The measurement of a chemical and/or its metabolites in human biological fluids or tissues. Also referred as to the internal dose of a chemical resulting from integrated exposures from all exposure routes.
Human data	They include observational studies (also called epidemiological studies) where the researcher is observing natural relationships between factors and health outcomes without acting upon study participants. Vigilance data also fall under this concept. In contrast, interventional studies (also called experimental studies or randomised clinical trials), where the researcher intercedes as part of the study design, are outside the scope of this opinion.
IARC	International Agency for Research on Cancer. An agency of the World Health Organization whose role is to conduct and coordinate research into the causes and occurrence of cancer worldwide.
LOAEL	Lowest-observed-adverse-effect level. The lowest concentration or amount of a chemical stressor evaluated in a toxicity test that shows harmful effects (e.g. an adverse alteration of morphology, biochemistry, function, or lifespan of a target organism).
NOAEL	No observed-adverse-effect level. Highest dose at which there was not an observed toxic or adverse effect.
OR	Odds ratio. A measure of association between an exposure and an outcome. The OR represents the odds that an outcome will occur given a particular exposure, compared to the odds of the outcome occurring in the absence of that exposure.
PBTK-TD	Physiologically based toxicokinetic/toxicodynamic modelling is a mathematical modelling approach aimed at integrating *a priori* knowledge of physiological processes with other known/observed information to mimic the fates and effects of compounds in the bodies of humans, preclinical species and/or other organisms.
PPP	Plant Protection Product. The term 'pesticide' is often used interchangeably with 'plant protection product', however, pesticide is a broader term that also covers non plant/crop uses, for example biocides.
RR	Relative risk. Ratio of the probability of an event (e.g. developing a disease) occurring in an exposed group to the probability of the event occurring in a comparison, non-exposed group.
RMS	Rapporteur member state. The member state of the European Union initially in charge of assessing and evaluating a dossier on a pesticide active substance toxicological assessment.
Sensitivity	The ability of a test to correctly classify an individual as 'diseased'. Probability of being test positive when disease present.
Specificity	The ability of a test to correctly classify an individual as disease-free. Probability of being test negative when disease absent.
Surrogate endpoint	A biomarker intended to substitute for a clinical endpoint

AHS	Agricultural Health Study
ASHTIII	Alerting and Reporting System for Chemical Health Threats, Phase III
BEES-C	Biomonitoring, Environmental Epidemiology, and Short-Lived Chemicals
DAR	draft assessment report
DDE	dichlorodiphenyldichloroethylene
DDT	dichlorodiphenyltrichloroethane
EMA	European Medicines Agency
EPA US	Environmental Protection Agency
EQUATOR	Enhancing the QUAlity and Transparency Of health Research
EU-OSHA	European Agency for Safety and Health at Work
EWAS	Exposome-wide association studies
GIS	Geographical information systems

GLP	good laboratory practice
GPS	global positioning system
HWE	healthy worker effect
IATA	Integrated Approach on Testing and Assessment
ICD	International Classification of Diseases
IHR	International Health Regulations
INSERM	French National Institute of Health and Medical Research
LOQ	limit of quantification
MGUS	monoclonal gammopathy of undetermined significance
MIE	molecular initiating event
MoA	mode of action
NHL	non-Hodgkin's lymphoma
NIOSH	National Institute for Occupational Safety and Health
NOS	Newcastle-Ottawa scale
OECD	Organisation for Economic Co-operation and Development
OPP	Office of Pesticide Programs
PCC	Poison Control Centre
PPE	personal protective equipment
RAR	Renewal Assessment Report
RASFF	rapid alert system covering food and feed
RTI	Research Triangle Institute
SAR	structure–activity relationship
STREGA	STROBE Extension to Genetic Association studies
STROBE	STrengthening the Reporting of OBservational studies in Epidemiology
ToR	Term of Reference
UF	uncertainty factor
WHO	World Health Organization
WoE	Weight-of-Evidence

Annex A – Pesticide epidemiological studies reviewed in the EFSA External Scientific Report and other reviews

The extensive evidence gathered by the EFSA External Scientific Report (Ntzani et al., 2013) highlights that there is a considerable amount of information available on pesticide exposure and health outcomes from epidemiological studies. Nonetheless, the quality of this evidence is usually low and many biases are likely to affect the results to an extent that firm conclusions cannot be made. In particular, exposure epidemiology has long suffered from poor measurement and definition and in particular for pesticides this has always been exceptionally difficult to assess and define.

A.1. The EFSA External scientific report

A.1.1. Methodological quality assessment

The External Scientific Report consists of a comprehensive systematic review of all the epidemiological studies published between 1 January 2006 and 30 September 2012, investigating the association between pesticide exposure and the occurrence of any human health-related outcomes.

The methodological assessment of eligible studies (to evaluate risk of bias associated with each study) was focused on: study design, study population, level of details in exposure definition and the methods of exposure measurement and the specificity of the measurement. Efforts undertaken to account for confounders through matching or multivariable models, blinded exposure assessment and well-defined and valid outcome assessment were considered.

The elements of the methodological appraisal were considered from the Research Triangle Institute (RTI; Research Triangle Park, NC, USA) item bank, a practical and validated tool for evaluating the risk of bias and precision of observational studies. Those elements are described below (Table A.1).

Table A.1: Elements from the Research Triangle Institute (RTI; Research Triangle Park, NC, USA) item bank for methodological appraisal of epidemiological studies

Question	High risk	Low risk
Study design (prospective, retrospective, mixed, NA)	Retrospective, mixed, NA	Prospective
Inclusion/exclusion criteria clearly stated (yes, partially, no)	No	Yes
Authors mention power calculations (yes, no)		Yes
Level of detail in describing exposure (high, medium, low)	Low	High
Robust measurement of exposure. (biomarker (yes); small area ecological measures, job titles, questionnaire (partial); was based on large area ecological measures (no)	No	Yes
Were measures of exposure specific? yes; based on broader, chemically-related groups (partial); based on broad groupings of diverse chemical and toxicological properties (no)	No	Yes
Attempt to balance the allocation between the groups (e.g., through stratification, matching)	No	Yes
Adjustment performed for potential confounders (yes, some, no)	No	Yes
Assessors blinded to exposure status (for cohort studies)	No	Yes
Outcomes assessed using valid and reliable measures, implemented consistently across all study participants?	No	Yes
Sample size	Low	Top
Rough quality assessment	>6 answers high risk	>6 asnwers low risk

Quantitative synthesis of the results was attempted when there were 5 or more eligible studies per examined outcome and when there was no substantial heterogeneity among the published evidence. Publication bias was assessed using funnel plots which allowed to visually inspect asymmetry when more than 10 studies were included in the meta-analysis.

Toxicological data was not reviewed or discussed in the External Scientific Report.

A.1.2. Inclusion/exclusion criteria

All types of pesticides, including those banned in the EU, were considered to enhance the totality of the epidemiological evidence available at the time of the review.

Exclusion criteria:

- Studies without control populations (case reports, case series) and ecological studies
- Pesticide poisoning or accidental high dose exposure
- Studies with no quantitative information on effect estimates
- Studies with different follow-up periods and examining the same outcome, only the one with the longest follow-up was retained to avoid data duplication.
- Studies referred to the adverse effects of substances used as therapy for various medical conditions (e.g. warfarin-based anticoagulants)
- Studies on solvents and other non-active ingredients (e.g. co-formulants) in pesticides
- Studies examining the association between exposure and biomarkers of exposure were not considered eligible as they do not examine health outcomes
- Studies/analyses investigating exposure to pesticides: arsenic, hexachlorocyclohexane (HCH) α or β, lead, dioxins and dioxin-like compounds including polychlorinated biphenyls (PCBs) were not considered
- Narrative reviews were excluded but not systematic reviews or meta-analyses.

Publications reporting series of acute poisonings or clinical cases, biomonitoring studies unrelated to health effects, or studies conducted on animals or human cell systems were not included; only epidemiological studies addressing human health effects were selected. Publications that lacked quantitative data for measuring associations were also excluded.

Cohort studies, case–control studies and cross-sectional studies were included. Each study underwent an assessment of its eligibility based on a method including 12 criteria such as study design, precise description of the inclusion/exclusion criteria, level of detail in describing exposure, robustness in the measurement of exposure, adjustment for potential confounding factors, method of assessment of the health outcome, sample size, etc. Among these 12 criteria, three were related to the degree of precision in the description/measurement of exposure, which may explain why a large number of epidemiological studies were not selected.

A.1.3. Results

Overall, 602 individual publications were included in the scientific review. These 602 publications corresponded to 6,479 different analyses. The overwhelming majority of evidence comes from retrospective or cross-sectional studies (38% and 32%, respectively) and only 30% of studies had a prospective design. Exposure assessment varied widely between studies and overall 46% measured biomarkers of pesticides exposure and another 46% used questionnaires to estimate exposure to pesticides. Almost half of the studies (49%) were based in America. Most studies examined associations between occupational exposure to pesticides and health effects. The entire spectrum of diseases associated with pesticides has not been studies before. The report examined a wide variety of outcomes (Figure A.1). The largest proportion of studies pertains to cancer outcomes (N = 164) and outcomes related to child health (N = 84).

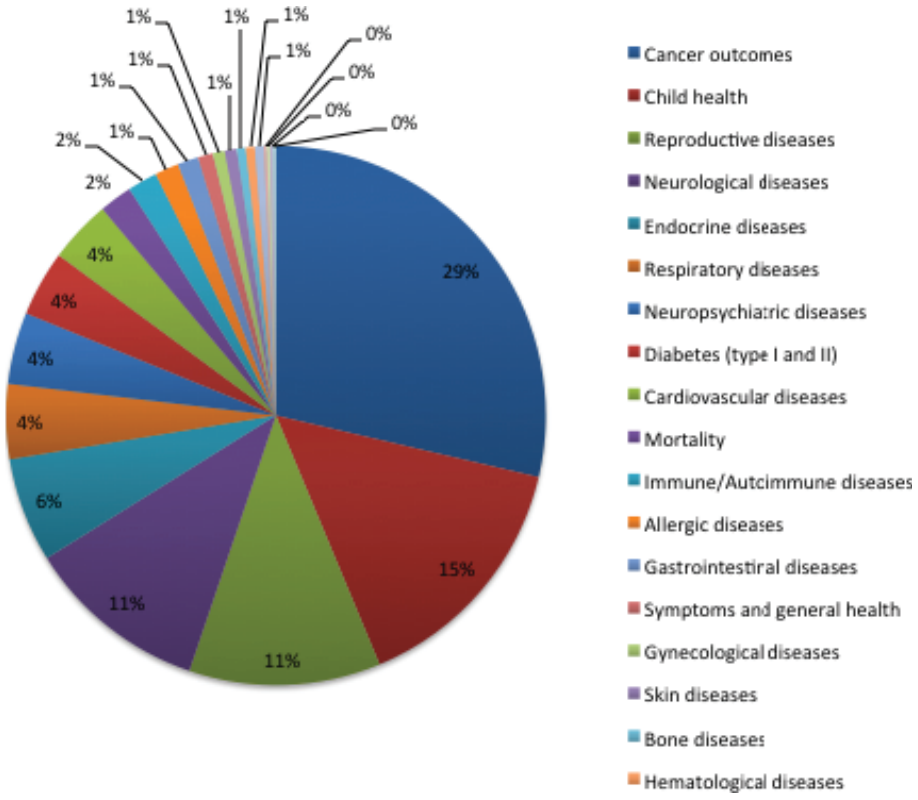


Figure A.1: Major outcome categories and corresponding percentage of studies examining those outcomes among the publications reviewed by the EFSA external scientific report (Ntzani et al., 2013)

Despite the large volume of available data and the large number (> 6,000) of analyses available, firm conclusions were not made for the majority of the outcomes studied. This was due to several limitations of the data collected as well as to inherent limitations of the review itself. As mentioned above, the review studied the whole range of outcomes examined in relation to pesticides during an approximately 5 years' period. Thus, only recent evidence was reviewed and the results of the meta-analyses performed should be cautiously interpreted as they do not include all the available evidence. It is therefore capable of highlighting outcomes which merit further in-depth analysis in relation to pesticides by looking at the entire literature (beyond 5 years) and by focusing on appraising the credibility of evidence selected. The limitations of the studies itself are in line with other field of environmental epidemiology and focus around the exposure assessment, the study design, the statistical analysis and reporting. In particular:

a) **Exposure assessment**: The assessment of exposure is perhaps the most important methodological limitation of the studies reviewed in the ESR. Studies used different methods for exposure assessment and assignment. Most studies were based on self-reported exposure to pesticides, defined as 'ever versus never' use or as 'regular versus non-regular' use. Such methods suffer from high misclassification rates and do not allow for dose–response analysis. This is especially the case for retrospective studies where misclassification would be differential with higher exposures reported in participants with disease (recall bias) (Raphael, 1987). While questionnaires might be capable of differentiating subjects with very high and very low exposure levels, they are not capable of valid exposure classification across an exposure gradient, thus not allowing the study of dose–response relationships. Also, questionnaire for exposure assessment need to be validated for use in epidemiological studies. Nonetheless, a vast proportion of studies use in house version of non-validated questionnaires which may suffer from content (the questionnaire does not cover all sources of exposure to the hazard of interest) or criterion validity (e.g. through inaccurate recall or misunderstanding of questions) (Coggon, 1995).

Although the range of categories of pesticide studied is wide, studies very often concentrate on a broadly defined pesticide category, so that it is difficult to know what type of pesticide the population is exposed to.

Exposure to pesticides was defined as reported use of pesticides by the study participant or by government registry data. These derive from self-administered questionnaires, interviewer administrated questionnaires, job exposure matrices (JEM), by residential status (proximity to pesticide exposure), by detecting biomarkers associated with pesticide exposure or by other means as defined by each study.

Studies often examine pesticides that have already been banned in western populations and the EU. The use of biomarkers as means of exposure assessment is infrequent, but still available in almost half of the studies.

b) **Study design**: As mentioned above, the majority of evidence comes form case–control studies and cross-sectional studies. Cross-sectional, and in part also case–control studies, cannot fully assess the temporal relationships and thus are less able to provide support regarding the causality of associations.

c) **Outcomes examined**: The definition of clinical outcomes displayed large variability in eligible epidemiological studies, which can further cause the variability in results. Perhaps most important in this setting is the use of a great number of surrogate outcomes examined. Surrogate outcomes are biomarkers or physical measures that are generally accepted as substitutes for, or predictors of, specific clinical outcomes. However, often these surrogate outcomes are not validated and do not meet the strict definitions of surrogate outcomes. Such outcomes can be defined as possible predictors of clinical outcomes but do not fulfil the criteria for a surrogate outcome. It is essential to appraise the evidence around non-validated surrogate outcomes by taking into account the implicit assumptions of these outcomes.

A great variety of assessed outcomes covering a wide range of pathophysiologies was observed. 'Hard' clinical outcomes as well as many surrogate outcomes included in the database reflect the different methodologies endorsed to approach the assessed clinical research questions. The different outcomes were divided into 23 major disease categories, with the largest proportion of studies addressing cancer and child health outcomes.

The adverse health effects assessed included:

a) major clinical outcomes, such as cancer, respiratory (allergy), reproductive (decreased fertility, birth defects) and neurodegenerative (Parkinson's disease);
b) clinical surrogate outcomes, e.g. neurodevelopmental impairment (assessed by neurocognitive scales);
c) laboratory surrogate outcomes (e.g. liver enzyme changes).

For many adverse health effects attributed to pesticide exposure, there exist contradictory or ambiguous studies. Whether this results from lack of consistency or real heterogeneity warrants further clarification.

d) **Statistical analysis**:

Simultaneous exposure to multiple agents (heavy metals, solvents, suspended particulate matter etc.) from different sources is common. It may introduce further bias in the results as all of them may produce adverse health outcomes. Thus, it is essential to account for confounding from exposure to multiple agents in order to delineate true associations but this has not been possible in the overwhelming majority of evidence assessed in the EFSA external scientific report.

In addition, the evidence collected and appraised in the EFSA external scientific report (Ntzani et al., 2013) is likely to suffer from selective reporting and multiple testing. The studies reported a very wide range of analyses; 602 publications resulted in 6,000 analyses. The amount of multiple hypothesis testing is enormous. These analyses need to be adjusted for multiple hypothesis testing else, otherwise the results suffer from high false positive rate. Even when studies present only one analysis, selective reporting is always a possibility as has been shown in other epidemiological fields as well. In addition, when interpreting results one should also take into account that, especially for certain outcomes (e.g. cancers), the majority of evidence comes from single study populations and the Agricultural Health Study in particular.

A.1.4. Conclusion of the EFSA External Scientific Report

Regardless of the limitations highlighted above, the External Scientific Report (Ntzani et al., 2013) showed consistent evidence of a link between exposure to pesticides and Parkinson's disease and childhood leukaemia, which was also supported by previous meta-analyses. In addition, an increased risk was also found for diverse health outcomes less well studied to date, such as liver cancer, breast cancer and type II diabetes. Effects on other outcomes, such as endocrine disorders, asthma and allergies, diabetes and obesity showed increased risks and should be explored further.

Childhood leukaemia and Parkinson's disease are the two outcomes for which a meta-analysis after 2006 was found consistently showing an increased risk associated with pesticide exposure. Nonetheless, the exposure needs to be better studied to disentangle the effect of specific pesticide classes or even individual pesticides. Significant summary estimates have also been reported for other outcomes (summarised in Table A.2). However, as they represent studies from 2006 onwards results should be regarded as suggestive of associations only and limitations especially regarding the heterogeneity of exposure should always been taken into consideration. Data synthesis and statistical tools should be applied to these data in relation to specific outcomes, after the update of the results to include publications before 2006, in order to quantify the amount of bias that could exist and isolate outcomes where the association with pesticides is well supported even when estimates of bias are taken into account. Similarly, outcomes where further evidence is needed to draw firm conclusions need to be highlighted.

Table A.2: Summary of meta-analyses performed in the report

Health outcome	N studies	Meta-analysis results	I^2
Leukaemia	6	1.26 (0.93; 1.71)	59.4%
Hodgkin lymphoma	7	1.29 (0.81–2.06)	81.6%
Childhood leukaemia (exposure to pesticides during pregnancy)	6	1.67 (1.25–2.23)	81.2%
Childhood leukaemia (exposure to insecticides during pregnancy)	5	1.55 (1.14–2.11)	65%
Childhood leukaemia (exposure to insecticides during pregnancy – update Turner, 2010)	9	1.69 (1.35–2.11)	49.8%
Childhood leukaemia (exposure to unspecified pesticides during pregnancy)	5	2.00 (1.73–2.30)	39.6%
Childhood leukaemia (exposure to unspecified pesticides during pregnancy – update Turner, 2010)	11	1.30 (1.06–1.26)	26.5%
Childhood leukaemia (exposure to pesticides during childhood)	7	1.27 (0.96–1.69)	61.1%
Childhood leukaemia (exposure to insecticides during childhood – update Turner, 2010)	8	1.51 (1.28–1.78)	0%
Childhood leukaemia (exposure to unspecified pesticides during childhood – update Turner, 2010)	11	1.36 (1.19–1.55)	0%
Breast cancer (DDE exposure)	5	1.13 (0.81–1.57)	0%
Breast cancer	11	1.24 (1.08–1.43)	0%
Testicular cancer (DDE exposure)	5	1.40 (0.82–2.39)	59.5%
Stomach cancer	6	1.79 (1.30–2.47)	0%
Liver cancer	5	2.50 (1.57–3.98)	25.4%
Cryptorchidism	8	1.19 (0.96–1.49)	23.9%
Cryptorchidism (DDT exposure)	4	1.47 (0.98–2.20)	51%
Hypospadias (general pesticide exposure)	6	1.01 (0.74–1.39)	71.5%
Hypospadias (exposure to specific pesticides)	9	1.00 (0.84–1.18)	65.9%
Abortion	6	1.52 (1.09–2.13)	63.1%
Parkinson's disease	26	1.49 (1.28–1.73)	54.6%
Parkinson's disease (DDT exposure)	5	1.01 (0.78–1.30)	0%
Parkinson's disease (paraquat exposure)	9	1.32 (1.09–1.60)	34.1%
Amyotrophic lateral sclerosis	6	1.58 (1.31–1.90)	10%
Asthma (DDT exposure)	5	1.29 (1.14–1.45)	0%
Asthma (paraquat exposure)	6	1.40 (0.95–2.06)	53.3%
Asthma (chlorpyrifos exposure)	5	1.03 (0.82–1.28)	0%
Type 1 diabetes (DDE exposure)	8	1.89 (1.25–2.86)	49%
Type 1 diabetes (DDT exposure)	6	1.76 (1.20–2.59)	76.3%
Type 2 diabetes (DDE exposure)	4	1.29 (1.13–1.48)	0%

N = number of studies considered for the meta-analysis; in the column of meta-analysis results, the numbers represent the statistical estimate for the size of effect (odds ratio (OR), or relative risk (RR)) with the corresponding 95% confidence interval (CI). I^2 represents the percentage of total variation across studies that is due to heterogeneity.

A.2. The INSERM report

In September 2013, the French National Institute of Health and Medical Research (INSERM) released a literature review carried out with a group of experts on the human health effects of exposure to pesticides.[22] Epidemiological or experimental data published in the scientific literature up to June 2012 were analysed. The report was accompanied by a summary outlining the literature analysis and highlighting the main findings and policy lines, as well as the recommendations.

The INSERM report is composed of four parts: (1) exposure assessment, with a detailed description of direct and indirect methods to assess exposure in epidemiological studies; (2) epidemiology, with an inventory and analysis of epidemiological studies available in the literature up to 2012, and a scoring system to assess the strength of presumed association; (3) toxicology, with a review of toxicological data (metabolism, mode of action and molecular pathway) of some substances and assessment of biological plausibility; (4) recommendations.

The vast majority of substances identified by the INSERM report as having a presumed moderate or strong association with the occurrence of health effects are chemicals that are now prohibited. This is mainly driven by the fact that the majority of the diseases examined are diseases of the elderly; therefore, the studies performed to date are based on persons who were old at the time of the study and exposed many years ago. By definition, it is not yet possible to investigate the potential long term effects of many of the more recent products.

These substances belong to the group of organochlorine insecticides, such as DDT or toxaphene, or insecticides with cholinesterase-inhibiting properties, such as terbufos or propoxur.

Of the seven approved active substances identified by the INSERM expert appraisal report (the herbicides 2,4-D, MCPA, mecoprop, glyphosate, the insecticide chlorpyrifos, and the foliar fungicides mancozeb and maneb), all had a presumed moderate or weak association with haematopoietic cancers. Two of them (the foliar fungicides mancozeb and maneb) had a presumed weak association with Parkinson's disease and two (chlorpyrifos and glyphosate) had a presumed association with developmental impairment identified as weak or moderate in the expert appraisal.

A.2.1. Description of methods to assess exposure in epidemiological studies

Different methods (direct and indirect) have been developed to assess exposure, such as biological or environmental monitoring data, ad hoc questionnaires, job- or crop-exposure matrices, analysis of professional calendars, sales data, land use data, etc. According to the authors, these various tools can be combined with each other but, to date none has been validated as a reference method for estimating exposure in the context of occupational pesticide exposure assessment.

A.2.2. Epidemiology

The group of experts from INSERM carried out an inventory and analysis of epidemiological studies available in the literature, examining the possible association between pesticide exposure and health outcomes: eight cancer sites (non-Hodgkin lymphoma, leukaemia, lymphoma, multiple myeloma, prostate, testis, brain, melanoma), three neurodegenerative diseases (Parkinson's disease, Alzheimer's disease, amyotrophic lateral sclerosis), cognitive or depressive disorders, effects on reproductive function (fertility, pregnancy and child development) and childhood cancers. These are health outcomes that have been identified in previous studies as potentially related to pesticide exposure.

Epidemiological studies addressing primarily farmers, pesticide applicators and workers of the pesticide manufacturing industries, as well as the general population when it was relevant, were selected.

The INSERM group of experts established a hierarchy in the relevance of the studies, placing the meta-analysis at the top, then the systematic review, then the cohort study, and finally, the case–control study. Based on this hierarchy, a scoring system was defined to assess the strength of presumption of the association between exposure and the occurrence of health outcomes from the analysis of the study results; for each disease or pathological condition investigated, this score may vary depending on the quality, type and number of available studies, as, for example:

(++): strong presumption: based on the results of a meta-analysis, or several cohort studies or at least one cohort study and two case–control studies, or more than two case–control studies;

[22] INSERM. Pesticides. Effets sur la santé. Collection expertise collective, Inserm, Paris, 2013.

(+): moderate presumption: based on the results of a cohort study or a nested case–control study or two case–control studies;

(±): weak presumption: based on the results of one case–control study. This synthesis takes the work beyond the status of a simple mapping exercise.

A.2.3. Toxicological data

Toxicological data that were considered in the literature review were mainly those regarding metabolism, mode of action and molecular pathways. None of the studies provided as part of the procedures for placing products on the market were considered except if they were published in the open literature.

When substances were clearly identified in the epidemiological studies, a scoring system was defined to assess the biological plausibility from the study results: coherence with pathophysiological data and occurrence of health outcome.

(++): hypothesis supported by 3 mechanisms of toxicity;

(+): hypothesis supported by at least one mechanism of toxicity.

A.2.4. Findings

The major results of the INSERM report are summarised in Tables A.3–A.6.

Table A.3: Statistically significant associations between occupational exposure to pesticides and health outcomes in adults (health outcomes that were analysed in the review)

Health outcome	Type of population with significant risk excess	Strength of presumption[a]
NHL	Farmers, operators, manufacturing plant personnel	++
Prostate cancer	Farmers, operators, manufacturing plant personnel	++
Multiple myeloma	Farmers, operators	++
Parkinson's disease	Occupational and non-occupational exposure	++
Leukaemia	Farmers, operators, manufacturing plant personnel	+
Alzheimer's disease	Farmers	+
Cognitive disorders[b]	Farmers	+
Fertility and fecundability disorders	Occupational exposure	+
Hodgkin lymphoma	Agricultural workers	±
Testicular cancer	Agricultural workers	±
Brain cancer (glioma, meningioma)	Agricultural workers	±
Melanoma	Agricultural workers	±
Amyotrophic lateral sclerosis	Farmers	±
Anxiety, depression[b]	Farmers, farmers with a history of acute poisoning, operators	±

(a): Scoring system: strong presumption (++), moderate presumption (+), weak presumption (±).
(b): Almost all pesticides were organophosphates.

Table A.4: Associations between occupational or home use exposure to pesticides and cancers or developmental impairment in children (health outcomes that were analysed in the review) (only statistically significant associations are shown)

Health outcome	Type of exposure and population with significant risk excess	Strength of presumption[(a)]
Leukaemia	Occupational exposure during pregnancy, prenatal exposure (residential)	++
Brain cancer	Occupational exposure during pregnancy	++
Congenital malformation	Occupational exposure during pregnancy; Residential exposure during pregnancy (agricultural area, home use)	++ +
Fetal death	Occupational exposure during pregnancy	+
Neurodevelopment	Residential exposure during pregnancy (agricultural area, home use, food)[(b)]; Occupational exposure during pregnancy	++ ±

(a): Scoring system: strong presumption (++), moderate presumption (+), weak presumption (±).
(b): Organophosphates.

Table A.5: Findings related to approved active substances: epidemiological assessment and biological plausibility

Active substance	Classification	Strength of presumption[(a)]	Biological plausibility[(b)]
Organophosphates ***Insecticide***			
Chlorpyrifos	Acute Tox cat 3	Leukaemia (+) Neurodevelopment (+) NHL (±)	Yes (++) Yes (++) Yes (++)
Dithiocarbamates ***Fungicide***			
Mancozeb/Maneb	Repro cat 2	Leukaemia (+) Melanoma (+) Parkinson's disease (in combination with paraquat) (±)	? ? Yes (+)
Phenoxy herbicides ***Herbicide***			
2,4-D MCPA Mecoprop	Acute Tox cat 4 Acute Tox cat 4 Acute Tox cat 4	NHL (+) NHL (±) NHL (±)	? ? ?
Aminophosphonate glycine ***Herbicide***			
Glyphosate		NHL (+) Fetal death (±)	? ?

(a): Scoring system: strong presumption (++), moderate presumption (+), weak presumption (±).
(b): Scoring system: (++): hypothesis supported by 3 different known mechanisms of toxicity, (+): hypothesis supported by at least one mechanism of toxicity.

Table A.6: Findings related to non-approved active substances: epidemiological assessment and biological plausibility

Active substance	Ban in the EU	IARC classification	Strength of presumption[a]	Biological plausibility[b]
Dieldrin	1978	3 or 2 (US-EPA)	NHL[c] (±) Prostate cancer (±) Parkinson's disease (±)	Yes (+) Yes (+) ?
DDT/DDE	1978	2B	NHL (++) Testicular cancer (+) Child growth (++) Neurodevelopment (±) Impaired sperm parameters (+)	Yes (+) ? ? ? ?
Chlordane	1978	2B	NHL (±) Leukaemia (+) Prostate cancer (±) Testicular cancer (+)	Yes (+) Yes (+) Yes (+) ?
Lindane (γ-HCH)	2002/2004/2006/2007	2B[d]	NHL (++) Leukaemia (+)	Yes (++) Yes (++)
β-HCH	2002/2004/2006/2007	2B[d]	Prostate cancer (±)	?
Toxaphene	2004	2B	NHL[c] (±) Leukaemia (+) Melanoma (+)	Yes (++) Yes (++) Yes (+)
Chlordecone	2004	2B	Cancer prostate (++) Impaired sperm parameters (+) Neurodevelopment (+)	Yes (+) ? ?
Heptachlor	1978	2B	Leukaemia (+)	Yes (+)
Endosulfan	2005	Not classified	?	Yes (+)
Hexachlorobenzene (HCB)	1978	2B	Child growth (+)	?
Terbufos	2003/2007		NHL (+) Leukaemia (+)	? ?
Diazinon	2008		NHL (+) Leukaemia (+)	? ?
Malathion	2008	3	NHL (++) Leukaemia (+) Neurodevelopment (+) Impaired sperm parameters (+)	Yes (+) Yes (+) ? ?
Fonofos	2003		NHL (±) Leukaemia (+) Prostate cancer (+)	? ? ?
Parathion	2002	3	Melanoma (+)	?
Coumaphos	Never notified and authorised in the EU		Prostate cancer (+)	?
Carbaryl	2008	3	NHL (±) Melanoma (+) Impaired sperm parameters (+)	? ? ?
Propoxur	2002		Neurodevelopment (+) Fetal growth (+)	? ?
Carbofuran	2008		NHL (±) Prostate cancer (+)	? ?
Butylate	2003		NHL (+) Prostate cancer (+)	? ?
EPTC	2003		Leukaemia (+)	?

Active substance	Ban in the EU	IARC classification	Strength of presumption[a]	Biological plausibility[b]
Atrazine	2005	3	NHL (±) Fetal growth (+)	Yes (+) ?
Cyanizine	2002/2007		NHL[c] (±)	?
Permethrin	2002	3	Prostate cancer (+)	Yes (+)
Fenvalerate	1998	Not classified	Impaired sperm parameters (+)	?
Methyl bromide	2010	3	Testicular cancer (+)	?
Dibromoethane	Banned	2A	Impaired sperm parameters (+)	?
Dibromochloropropane (DBCP)	Banned	2B	Impaired sperm parameters/impaired fertility (+++) (causal association)	Yes (+++) (mode of action elucidated)
Paraquat	2007		Parkinson's disease (+)	Yes (++)
Rotenone	2011		Parkinson's disease (+)	Yes (++)
Alachlor	2008		Leukaemia (+)	Yes (++)

(a): Scoring system: strong presumption (++), moderate presumption (+), weak presumption (±).
(b): Scoring system: (++): hypothesis supported by 3 mechanisms of toxicity, (+): hypothesis supported by at least one mechanism of toxicity.
(c): Population with t(14,18) translocation, only.
(d): Technical mixture (α-, β-, and γ-HCH).

A.2.5. Recommendations

The analysis of the available epidemiological and mechanistic data on some active substances suggests several recommendations for developing further research:

a) Knowledge on population exposure to pesticides should be improved

1) Collect information about use of active substances by farmers
2) Conduct field studies to measure actual levels of exposure
3) Monitor exposure during the full occupational life span
4) Measure exposure levels in air (outdoor and indoor), water, food, soil
5) Collect information on acute poisonings
6) Improve analytical methods for biomonitoring and external measurements
7) Allow researchers to have access to extensive formulation data (solvents, co-formulants, etc.).

b) Research potential links between exposure and health outcomes

1) Characterise substances or groups of substances causing health outcomes
2) Focus on susceptible individuals or groups of individuals (gene polymorphism of enzymes, etc.)
3) Focus on exposure windows and susceptibility (pregnancy, development)
4) Bridge the gap between epidemiology and toxicology (mode of action)
5) Improve knowledge on mixture toxicity
6) Foster new approaches of research (*in vitro* and *in silico* models, omics, etc.).

A.3. Similarities and differences between the EFSA External Scientific Report and the INSERM report

The two reports discussed herein have used different methodologies. Yet, their results and conclusions in many cases agree. The INSERM report is limited to predefined outcomes and it attempted to investigate the biological plausibility of epidemiological studies by reviewing toxicological data as well, meanwhile the EFSA report is a comprehensive systematic review of all available epidemiological studies that were published during an approximately 5 year window.

The differences between the reports are shown in Table A.7 and are related to the time period of search (i.e. both reports did not assess the same body of published data), different criteria for eligibility of studies and different approaches to summarising the evidence across and within outcomes.

Overall, the INSERM report identified a greater number of associations with adverse health effects than the EFSA report. However, a well-documented association with pesticide exposure was claimed by both reports for the same health outcomes (childhood leukaemia, Parkinson's disease).

Table A.7: Comparison between methods used in the EFSA External Scientific Report and the INSERM Report

	EFSA External report	INSERM report
Articles reviewed	602/43,000	NR
Language	Yes	NR
Search strategy (key words, MeSH)	Yes	NR
Search database	Yes (4)	NR
Years of publication	2006–2012 (Sep)	? to 2012 (Jun)
Type of epi studies assessed	Cross-sectional	Cross-sectional
	Case–control	Case–control
	Cohort	Cohort
Inclusion criteria	Yes	NR
Exclusion criteria	Yes	NR
Methodological quality assessment	Yes (12 criteria)	NR
Exposure groups[(a)]	Yes	Yes
Exposure assessment	Yes	Yes
Quantitative synthesis (meta-analysis)	Yes	No
Qualitative synthesis[(c)]	Yes	Yes
Supporting Toxicological data	NI	Yes
Associations with individual pesticides	Yes	Yes
Health outcomes studied		
Haematological cancer	Yes	Yes
Solid tumours	Yes	Yes
Childhood cancer	Yes	Yes
Neurodegenerative disorders	Yes	Yes
Neurodevelopmental outcomes	Yes	Yes
Neuropsychiatric disturbances[(b)]	No	Yes
Reproductive and developmental	Yes	Yes
Endocrine	Yes	NI
Metabolism	Yes	Yes
Immunological	Yes	NI
Respiratory	Yes	NI

NR: not reported; NI: not investigated.
(a): Exposure type (environmental, occupational, etc.) and period (general population, children, etc.).
(b): E.g. depressive disorders.
(c): Add explanation.

A.4. The Ontario College of Family Physicians Literature review (OCFPLR)

In 2004, the Ontario College of Family Physicians (Ontario, Canada) reviewed the literature published between 1992 and 2003 on major health effects associated with pesticide exposure. The authors concluded that positive associations exist between solid tumours and pesticide exposures as shown in Table A.8. They noted that in large well-designed cohort studies these associations were consistently statistically significant, and the relationships were most consistent for high exposure levels. They also noted that dose–response relationships were often observed, and they considered the quality of studies to be generally good.

Table A.8: Health Effects considered in the Ontario College of Family Physicians review, 2004

Endpoint	Associations identified by the Ontario College, pesticide (if differentiated), study type, (no. of studies/total no. of studies)
A) Cancer	
1. Lung	−ve cohort (1/1) +ve case–control (1/1) +ve carbamate, phenoxy acid, case–control (1/1)
2. Breast	+ve case–control (2/4) +ve ecological (1/1) +ve triazine, ecological (1/1) −ve atrazine, ecological (1/1)
3. Colorectal	
4. Pancreas	+ve cohort (1/1) +ve case–control (2/2)
5. Non-Hodgkin's lymphoma	+ve cohort (9/11) +ve case–control (12/14) +ve ecological (2/2)
6. Leukaemia	+ve cohort (5/6) +ve case–control (8/8) −ve ecological (1/1) +ve lab study (1/1)
7. Brain	+ve cohort (5), similar case–control (5)
8. Prostate	+ve cohort (5/5) case–control (2/2) ecological (1/1)
9. Stomach	
10. Ovary	
11. Kidney	+ve pentachlorophenol cohort (1/1) +ve cohort (1/1) +ve case–control (4/4)
12. Testicular	
B) Non-Cancer	
1) Reproductive effects	+ve glyphosate
Congenital malformations	+ve pyridyl derivatives
Fecundity/time to pregnancy	Suggest impaired
Fertility	
Altered growth	Possible +ve association, but further study required
Fetal death	Suggested association
Mixed outcomes	
2) Genotoxic/immunotoxic Chromosome aberrations	+ve Synthetic pyrethroids (1) +ve organophosphates (1) +ve fumigant and insecticide applicators
NHL rearrangements	+ve fumigant and herbicide applicators
3) Dermatologic	
4) Neurotoxic Mental & emotional impact	+ve
Functional nervous system impact	+ ve organophosphate/carbamate poisoning
Neurodegenerative impacts (PD)	+ve cohort (4/4) +ve case–control (2/2) +ve ecological (1/1)

+ve: positive; −ve: negative.

The report concluded that there was compelling evidence of a link between pesticide exposure and the development of non-Hodgkin's lymphoma (NHL), and also clear evidence of a positive association between pesticide exposure and leukaemia. The authors also claimed to have found consistent findings of a number of nervous system effects, arising from a range of exposure time courses.

Such strong conclusions found favour with Non-Governmental organisations (NGOs) and raised questions among some Regulatory Authorities. The Advisory Committee on Pesticides (ACP), at that time an UK government independent advisory committee, was asked to provide an evaluation of the outcome of the Ontario College review. The committee membership included one epidemiologist and the committee consulted five other epidemiologists involved in providing independent advice to other government committees. They all agreed that the review had major shortcomings (e.g. exact search strategy and selection criteria not specified, selective reporting of results, inadequate understanding and consideration of relevant toxicology, insufficient attention to routes and levels of exposure, not justified conclusions, etc.). Overall, the conclusions of the Ontario College review were considered not to be supported by the analysis presented. In 2012, the Ontario review authors published an update of their evaluation; in their second report they used a very similar approach but offered more detail concerning the inclusion criteria used. This example is a reminder of the risk of over interpretation of epidemiological studies. In particular, a causal inference between exposure and the occurrence of adverse health effects is often made, but this represents an association that should be further assessed.

Annex B – Human biomonitoring project outsourced by EFSA[23]

In 2015, EFSA outsourced a project to further investigate the role of HBM in occupational health and safety strategies as a tool for refined exposure assessment in epidemiological studies and to contribute to the evaluation of potential health risks from occupational exposure to pesticides. It was in fact recognised that exposure assessment is a key part of all epidemiological studies and misclassification of exposure and use of simple categorical methods are known to weaken the ability of a study to determine whether an association between contact and ill-health outcome exists; at present, this limits integration of epidemiological findings into regulatory risk assessment.

The consortium formed by Risk & Policy Analysts Limited (RPA), IEH Consulting Limited (IEH) and the Health&Safety Laboratory (HSL) carried out a systematic literature review for the period 1990–2015 with the aim to provide an overview on the use of HBM as a tool for occupational exposure assessment refinement, identifying advantages, disadvantages and needs for further development (first objective). The search identified 2096 publications relating to the use of HBM to assess occupational exposure to pesticides (or metabolites). The outcome of the search (Bevan et al., 2017) indicated that over the past 10–20 years there has been an expansion in the use of HBM, especially into the field of environmental and consumer exposure analysis. However, further improvement of the use of HBM for pesticide exposure assessment is needed, in particular with regards to: development of strategies to improve or standardise analytical quality, improvement of the availability of reference material for metabolites, integration of HBM data into mathematical modelling, exposure reconstruction, improvements in analytical instrumentation and increased availability of human toxicology data.

The contractors performed a review of available HBM studies/surveillance programmes conducted in EU/US occupational settings to identify pesticides (or metabolites) both persistent and not persistent, for which biomarkers of exposure (and possibly effect) were available and validated (second objective). A two-tiered screening process that included quality scoring for HBM, epidemiological and toxicological aspects, was utilised to identify the most relevant studies, resulting in 178 studies for critical review. In parallel with the screening of identified studies, a Master Spreadsheet was designed to collate data from these papers, which contained information relating to: study type; study participants; chemicals under investigation; biomarker quality check; analytical methodology; exposure assessment; health outcome/toxicological endpoint; period of follow-up; narrative of results; risk of bias and other comments.

HBM has been extensively used for monitoring worker exposure to a variety of pesticides. Epidemiological studies of occupational pesticide use were seen to be limited by inadequate or retrospective exposure information, typically obtained through self-reported questionnaires, which can potentially lead to exposure misclassification. Some examples of the use of job exposure or crop exposure matrices were reported. However, little validation of these matrix studies against actual exposure data had been carried out. Very limited data was identified that examined seasonal exposures and the impact of PPE, and many of the studies used HBM to only assess one or two specific compounds. A wide variety of exposure models are currently employed for health risk assessments and biomarkers have also often been used to evaluate exposure estimates predicted by a model.

From the 178 publications identified to be of relevance, 41 individual studies included herbicides, and of these, 34 separate herbicides were identified, 15 of which currently have approved for use in the EU. Similarly, of the 90 individual studies that included insecticides, 79 separate insecticides were identified, of which 18 currently have approved for use in the EU. Twenty individual studies included fungicides, with 34 separate fungicides being identified and of these 22 currently have approved for use in the EU. The most studied herbicides (in order) were shown to be: 2,4-D > atrazine > metolachlor = MCPA > alachlor = glyphosate. Similarly, the most studied insecticides (in order) were: chlorpyrifos > permethrin > cypermethrin = deltamethrin > malathion, and the most studied fungicides were: captan > mancozeb > folpet.

Current limitations comprised the limited number of kinetic data from humans, particularly with respect to the ADME of individual pesticides in human subjects, which would allow more accurate HBM sampling for all routes of exposure. A wider impact of this is on the development of PBPK models for the risk assessment of pesticides, which rely on toxicokinetic data, and on validation of currently used exposure assessment models. Further limitations currently impacting on the use of HBM in this field are a lack of large prospective cohort studies to assess long term exposure to currently used pesticides.

[23] Bevan et al. (2017).

The evidence identified has been used to help formulate recommendations on the implementation of HBM as part of the occupational health surveillance for pesticides in Europe. Some key issues were considered that would need to be overcome to enable implementation. These included the setting of priorities for the development of new specific and sensitive biomarkers, the derivation and adoption of health-based guidance values, development of QA schemes to validate inter-laboratory measurements, good practice in field work and questionnaire design, extension of the use of biobanking and the use of HBM for post-approval monitoring of pesticide safety.

Annex C – Experience of international regulatory agencies in regards to the integration of epidemiological studies for hazard identification

C.1. WHO-International Agency for Research on Cancer (IARC)

The IARC Monographs on the Evaluation of Carcinogenic Risks to Humans of the International Agency for Research on Cancer (IARC) is a programme established four decades ago to assess environmental exposures that can increase the risk of human cancer. These include individual chemicals and chemical mixtures, occupational exposures, physical agents, biological agents and lifestyle factors.

IARC assembles international interdisciplinary Working Groups of scientists to review and assess the quality and strength of evidence from scientific publications and perform a hazard evaluation to assess the likelihood that the agents of concern pose a cancer risk to humans. In particular, the tasks of IARC Working Group Members include the evaluation of the results of epidemiological and other experimental studies on cancer, to evaluate data on the mechanisms of carcinogenesis and to make an overall evaluation of the carcinogenicity of the exposure to humans.

The Monographs are widely used and referenced by governments, organisations, and the public around the world to set preventive and control public health measures.

The Preamble[24] to the IARC Monographs explains the scope of the programme, the scientific principles and procedures used in developing a Monograph, the types of evidence considered and the scientific criteria that guide the evaluations. The scope of the monographs broadened to include not only single chemicals but also groups of related chemicals, complex mixtures, occupational exposures, physical and biological agents and lifestyle factors. Thus, the title of the monographs reads 'Evaluation of carcinogenic risks to humans'.

Relevant epidemiological studies, cancer bioassays in experimental animals, mechanistic data, as well as exposure data are critically reviewed. Only reports that have been published or accepted for publication in the openly available scientific literature are included. However, the inclusion of a study does not imply acceptance of the adequacy of the study design or of the analysis and interpretation of the results. Qualitative aspects of the available studies are carefully scrutinised.

Although the Monographs have emphasised hazard identification, the same epidemiological and experimental studies used to evaluate a cancer hazard can also be used to estimate a dose–response relationship. A Monograph may undertake to estimate dose–response relationships within the range of the available epidemiological data, or it may compare the dose–response information from experimental and epidemiological studies.

The structure of a Monograph includes the following sections:

1) Exposure data
2) Studies of cancer in humans
3) Studies of cancer in experimental animals
4) Mechanistic and other relevant data
5) Summary
6) Evaluation and rationale.

Human epidemiological data are addressed in point 2, where all pertinent epidemiological studies are assessed. Studies of biomarkers are included when they are relevant to an evaluation of carcinogenicity to humans.

The IARC evaluation of epidemiological studies includes an assessment of the following criteria: types of studies considered (e.g. cohort studies, case–control studies, correlation (or ecological) studies and intervention studies, case reports), quality of the study (e.g. bias, confounding, biological variability and the influence of sample size on the precision of estimates of effect), meta analysis and pooled analyses, temporal effects (e.g. temporal variables, such as age at first exposure, time since first exposure, duration of exposure, cumulative exposure, peak exposure), use of biomarkers in epidemiological studies (e.g. evidence of exposure, of early effects, of cellular, tissue or organism responses), and criteria for causality.

With specific reference to causality, a judgement is made concerning the strength of evidence that the agent in question is carcinogenic to humans. In making its judgement, the Working Group

[24] http://monographs.iarc.fr/ENG/Preamble/CurrentPreamble.pdf

considers several criteria for causality (Hill, 1965). A strong association (e.g. a large relative risk) is more likely to indicate causality. However, it is recognised that weak associations may be important when the disease or exposure is common. Associations that are replicated in several studies of different design under different exposure conditions are more likely to represent a causal relationship than isolated observations from single studies. In case of inconsistent results among different investigations, possible reasons (e.g. differences in exposure) are sought, and high quality studies are given more weight compared to less methodologically sound ones. Risk increasing with the exposure is considered to be a strong indication of causality, although the absence of a clear dose–response effect is not necessarily evidence against a causal relationship. The demonstration of a decline in risk after cessation of or reduction in exposure also supports a causal interpretation of the findings. Temporality, precision of estimates of effect, biological plausibility and coherence of the overall data are considered. Biomarkers information may be used in an assessment of the biological plausibility of epidemiological observations. Randomised trials showing different rates of cancer among exposed and unexposed individuals provide particularly strong evidence for causality.

When epidemiological studies show little or no indication of an association between an exposure and cancer, a judgement of lack of carcinogenicity can be made. In those cases, studies are scrutinised to assess the standards of design and analysis described above, including the possibility of bias, confounding or misclassification of exposure. In addition, methodologically sound studies should be consistent with an estimate of effect of unity for any observed level of exposure, provide a pooled estimate of relative risk near to unity, and have a narrow confidence interval. Moreover, no individual study nor the pooled results of all the studies should show any increasing risk with increasing level of exposure. Evidence of lack of carcinogenicity can apply only to the type(s) of cancer studied, to the dose levels reported, and to the intervals between first exposure and disease onset observed in these studies. Experience with human cancer indicates that the period from first exposure to the development of clinical cancer is sometimes longer than 20 years, and latent periods substantially shorter than 30 years cannot provide evidence for lack of carcinogenicity.

Finally, the body of evidence is considered as a whole, in order to reach an overall evaluation which summarises the results of epidemiological studies, the target organs or tissues, dose–response associations, evaluations of the strength of the evidence for human and animal data, and the strength of the mechanistic evidence.

At the end of the overall evaluation, the agent is assigned to one of the following groups: Group 1, the agent is carcinogenic to humans; Group 2A, the agent is probably carcinogenic to humans; Group 2B, the agent is possibly carcinogenic to humans; Group 3, the agent is not classifiable as to its carcinogenicity to humans; Group 4, the agent is probably not carcinogenic to humans.

The categorisation of an agent is a matter of scientific judgement that reflects the strength of the evidence derived from studies in humans and in experimental animals and from mechanistic and other relevant data. These categories refer only to the strength of the evidence that an exposure is carcinogenic and not to the extent of its carcinogenic activity (potency).

For example, Group 1: The agent is carcinogenic to humans. This category is used when there is sufficient evidence of carcinogenicity in humans. Exceptionally, an agent may be placed in this category when evidence of carcinogenicity in humans is less than sufficient but there is sufficient evidence of carcinogenicity in experimental animals and strong evidence in exposed humans that the agent acts through a relevant mechanism of carcinogenicity.

Although widely accepted internationally, there have been criticisms of the classification of particular agents in the past, and more recent criticisms have been directed at the general approach adopted by IARC for such evaluations possibly motivating publication of a rebuttal (Pearce et al., 2015).

C.2. The experience of US-EPA in regards to the integration of epidemiological studies in risk assessment

The US Environmental Protection Agency's Office of Pesticide Programs (OPP) is the governmental organisation in the US responsible for registering and regulating pesticide products.[25] As part of this activity and prior to any permitted use of a pesticide, OPP evaluates the effects of pesticides on human health and the environment. EPA receives extensive hazard and exposure information to characterise

[25] See https://www.epa.gov/pesticide-science-and-assessing-pesticide-risks for general information on pesticide science and assessing pesticide risks.

the risks of pesticide products through the Federal Insecticide, Fungicide, and Rodenticide Act (FIFRA) and the Federal Food, Drug, and Cosmetic Act (FFDCA). Information on the toxic effects of pesticides is generally derived from studies with laboratory animals conducted by pesticide registrants and submitted to EPA.

In the past, information from well-designed epidemiology studies on pesticides has not been typically available to inform EPA's evaluations of potential risks that might be associated with exposure to pesticides. With an increasing number of epidemiology studies entering the literature which explore the putative associations between pesticides exposure and health outcomes, EPA is putting additional emphases on this source of information. This is especially true for the wealth of studies deriving from the Agricultural Health Study[26] (AHS), a large, well-conducted prospective cohort study following close to 90,000 individuals over more than 20 years and from the Children's Environmental Health and Disease Prevention Research Centers.[27] EPA intends to make increasing use of these epidemiology studies in its human health risk assessment with the goal of using such epidemiological information in the most scientifically robust and transparent way.

C.2.1. OPP Epidemiological Framework Document

As an early first step in this process, EPA-OPP developed a proposed epidemiological framework document released as a draft in 2010, 'Framework for Incorporating Human Epidemiologic and Incident Data in Health Risk Assessment' (US-EPA, 2010a). The 2010 draft framework was reviewed favourably by the FIFRA Scientific Advisory Panel (SAP) in February, 2010 (US-EPA, 2010b). This document was recently updated in 2016 to the 'Office of Pesticide Programs' Framework Document for Incorporating Human Epidemiology and Incident Data in Risk Assessments for Pesticides' (US-EPA, 2016). The revised and updated 2016 Framework document proposes that human information like that found in epidemiology studies (in addition to human incident databases, and biomonitoring studies) along with experimental toxicological information play a significant role in this new approach by providing insight into the effects caused by actual chemical exposures. In addition, epidemiological/ molecular epidemiological data can guide additional analyses, identify potentially susceptible populations and new health effects and potentially confirming existing toxicological observations. The concepts in the 2016 Framework are based on peer-reviewed robust principles and tools and rely on many existing guidance documents and frameworks (Table C.1) for reviewing and evaluating epidemiology data. It is also consistent with updates to the World Health Organization/International Programme on Chemical Safety mode of action (MoA)/human relevance framework which highlight the importance of problem formulation and the need to integrate information at different levels of biological organisation (Meek et al., 2014). Furthermore, it is consistent with recommendations by the National Academy of Sciences' National Research Council (NAS/NRC) in its 2009 report *Science and Decisions* (NRC, 2009) in that the framework describes the importance of using problem formulation at the beginning of a complex scientific analysis. The problem formulation stage is envisioned as starting with a planning dialogue with risk managers to identify goals for the analysis and possible risk management strategies. This initial dialogue provides the regulatory context for the scientific analysis and helps define the scope of such an analysis. The problem formulation stage also involves consideration of the available information regarding the pesticide use/usage, toxicological effects of concern, exposure pathways, and duration along with key gaps in data or scientific information.

[26] See https://aghealth.nih.gov/
[27] See https://www.epa.gov/research-grants/niehsepa-childrens-environmental-health-and-disease-prevention-research-centers

Table C.1: Key guidance documents and frameworks used by OPP (from US-EPA, 2016)

	1983	Risk Assessment in the Federal Government. Managing the Process
NAS	1994	Science and Judgement
	2007	Toxicity testing in the 21st Century
	2009	Science and Decisions: Advancing Risk Assessment
WHO/ IPCS	2001–2007	Mode of Action/Human Relevance Framework
	2005	Chemical Specific Adjustment Factors (CSAF)
	2014	New Development in the evolution and application of the WHO/IPCS framework on mode of action/species concordance analysis
EPA	1991–2005	Risk Assessment Forum Guidance for Risk Assessment (e.g. guidelines for carcinogen, reproductive, developmental, neurotoxicity, ecological, and exposure assessment, guidance for benchmark dose modelling, review of reference dose and reference concentration processes) http://www.epa.gov/risk_assessment/guidance.htm
	2000	Science Policy Handbook on Risk Characterisation http://nepis.epa.gov/Exe/ZyPURL.cgi?Dockey=40000006.txt
	2006	Approaches for the Application of Physiologically Based Pharmacokinetic (PBPK) Models and Supporting Data for Risk Assessment
	2014	Framework for Human Health Risk Assessment to Inform Decision-making
	2014	Guidance for Applying Quantitative Data to Develop Data-Derived Extrapolation Factors for Inter-species and Intra-species Extrapolation
	2001	Aggregate Risk Assessment https://www.epa.gov/sites/production/files/2015-07/documents/aggregate.pdf
OPP	2001 and 2002	Cumulative Risk Assessment http://www.epa.gov/ncer/cra/
OECD	2013	Organisation for Economic Co-operation and Development Guidance Document on Developing and Assessing Adverse Outcome Pathways

Briefly, this EPA Framework document describes the scientific considerations that the Agency will weigh in evaluating how such epidemiological studies and scientific information can be integrated into risk assessments of pesticide chemicals and also in providing the foundation for evaluating multiple lines of scientific evidence in the context of the understanding of the adverse outcome pathway (or MoA). The framework relies on and espouses standard practices in epidemiology, toxicology and risk assessment, but allows for the flexibility to incorporate information from new or additional sources. One of the key components of the Agency's framework is the use the MoA framework/adverse outcome pathway concept as a tool for organising and integrating information from different sources to inform the causal nature of links observed in both experimental and observational studies. MoA (Boobis et al., 2008; Simon et al., 2014; Meek et al., 2014) and adverse outcome pathway (Ankley et al., 2010) provide important concepts in the integrative analysis discussed in the Framework document. Both a MoA and an adverse outcome pathway are based on the premise that an adverse effect caused by exposure to a compound can be described by a series of causally linked biological key events that result in an adverse human health outcome, and have as their goal a determination of how exposure to environmental agents can perturb these pathways, thereby causing a cascade of subsequent key events leading to adverse health effects.

A number of concepts in the Framework are taken from two reports from the National Academies, *Science and Decisions: Advancing Risk Assessment* (NAS 2009) and *Toxicity Testing on the 21st Century* (NAS 2007). These two NRC reports advocate substantial changes in how toxicity testing is performed, how such data are interpreted, and ultimately how regulatory decisions are made. In particular, the 2007 report on 21st century toxicity testing advocates a decided shift away from the current focus of using apical toxicity endpoints to using toxicity pathways to better inform toxicity testing, risk assessment, and decision-making.

The MoA framework begins with the identification of the series of key events that are along the causal path and established on weight of evidence using criteria based on those described by Bradford Hill taking into account factors such as dose–response, temporal concordance, biological plausibility, coherence and consistency. Specifically, the modified Bradford Hill Criteria (Hill, 1965) are used to evaluate the experimental support that establishes key events within a MoA or an adverse outcome pathway, and explicitly considers such concepts as strength, consistency, dose response, temporal

concordance, and biological plausibility in a weight of evidence analysis. Using this analytic approach, epidemiological findings can be evaluated in the context of other human information and experimental studies to evaluate consistency, reproducibility, and biological plausibility of reported outcomes and to identify areas of uncertainty and future research. Figure C.1 below (adapted from NRC, 2007) suggests how different types of information relate to each other across multiple levels of biological organisation (ranging from the molecular level up to population-based surveillance) and is based on the rapidly evolving scientific understanding of how genes, proteins, and small molecules interact to form molecular pathways that maintain cell function in humans.

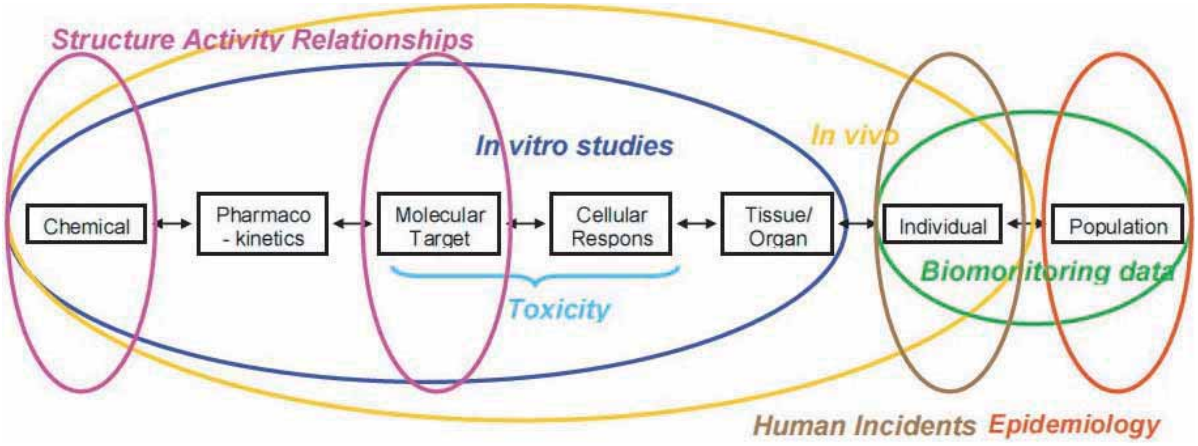


Figure C.1: Source to Outcome Pathway: Chemical effects across levels of biological organisation (adapted from NRC, 2007)

C.2.2. Systematic reviews: Fit for purpose

The National Academies' National Research Council (NRC) in its review of EPA's IRIS program defines systematic review as 'a scientific investigation that focuses on a specific question and uses explicit, prespecified scientific methods to identify, select, assess, and summarise the findings of similar but separate studies'.[28] In recent years, the NRC has encouraged the agency to move towards systematic review processes to enhance the transparency of scientific literature reviews that support chemical-specific risk assessments to inform regulatory decision-making.[29]

Consistent with NRC's recommendations, EPA-OPP employs fit-for-purpose systematic reviews that rely on transparent methods for collecting, evaluating and integrating the scientific data supporting its decisions. As such, the complexity and scope of each systematic review will vary among risk assessments. EPA-OPP starts with scoping/problem formulation followed by data collection, data evaluation, data integration and summary findings with critical data gaps identified.

Systematic reviews often use statistical (e.g. meta-analysis) and other quantitative techniques to combine results of the eligible studies, and can use a semi-quantitative scoring system to evaluate the levels of evidence available or the degree of bias that might be present. For EPA's Office of Pesticide Programs, such a Tier III (systematic review) assessment conducted as part of its regulatory review process would involve review of the pesticide chemical undergoing review and a specific associated suspected health outcome (as suggested by the initial Tier II assessment).

A number of federal and other organisations in the US are evaluating or have issued guidance documents for methods to conduct such systematic reviews and a number of frameworks have been

[28] http://dels.nas.edu/Report/Review-Integrated-Risk/18764

[29] NRC, 2011. 'Review of the Environmental Protection Agency's Draft IRIS Assessment of Formaldehyde' available for download at https://www.nap.edu/catalog/13142/review-of-the-environmental-protection-agencys-draft-iris-assessment-of-formaldehyde; See also NRC, 2014. 'Review of EPA's Integrated Risk Information System (IRIS) Process' available for download at https://www.nap.edu/catalog/18764/review-of-epas-integrated-risk-information-system-iris-process

developed. These include the EPA IRIS programs' approach,[30] the National Toxicology Programs' Office of Health Assessment and Translation (NTP/OHAT) approach[31] the Cochran Collaboration's approach,[32] the Campbell Collaboration and the Navigation Guide,[33] with this latter described in a series of articles in the journal *Environmental Health Perspectives*. Each broadly shares four defined steps: data collection, data evaluation, data integration, and summary/update. For example, The Cochrane Collaboration in its Cochrane Handbook for Systematic Reviews of Interventions for evidence-based medicine lists a number of the important key characteristics of a systematic review to be (from US-EPA, 2016):

- a clearly stated set of objectives with predefined eligibility criteria for studies;
- an explicit, reproducible methodology;
- a systematic search that attempts to identify all studies that would meet the eligibility criteria;
- an assessment of the validity of the findings from the identified studies;
- a systematic presentation and synthesis of the characteristics and findings of the included studies.

As described and elaborated in the following sections of this Annex, OPP's approach to review and integration of epidemiological data into pesticide risk assessments takes a tiered approach which each tier appropriately fit-for-purpose in the sense that is considers 'the usefulness of the assessment for its intended purpose, to ensure that the assessment produced is suitable and useful for informing the needed decisions (US-EPA, 2012) and that required resources are matched or balanced against any projected or anticipated information gain from further more in-depth research. A Tier 1 assessment is either a scoping exercise or an update to a scoping exercise in which a research and evaluation is limited to studies derived from the AHS. A Tier II assessment involves a broader search of the epidemiological literature, comprehensive data collection, and a deeper, more involved data evaluation and is more extensive but is generally limited in scope to epidemiology and stops short of multidisciplinary integration across epidemiology, human poisoning events, animal toxicology and adverse outcome pathways. A Tier III assessment is a complete systematic review with data integration and more extensive data evaluation and extraction and may involve more sophisticated epidemiological methods such as meta-analysis and meta-regression, causal inference/causal diagrams, and quantitative bias and sensitivity analyses, among others.

C.2.3. Current and Anticipated Future EPA Epidemiology Review Practices

C.2.3.1. Tier I (Scoping & Problem Formulation) and Tier II (more extensive literature search)

Currently at EPA, epidemiology review of pesticides is conducted in a tiered process as the risk assessment develops, as briefly described above. The purpose of this early Tier I/scoping epidemiology report is to ensure that highly relevant epidemiology studies are considered in the problem formulation/scoping phase of the process and, if appropriate, fully reviewed in the (later) risk assessment phase of the process. In Tier I, EPA-OPP focuses on well-known high quality cohort studies which focus on pesticide issues, particularly the Agricultural Health Study (AHS). The AHS is a federally funded study that evaluates associations between pesticide exposures and cancer and other health outcomes and represents a collaborative effort between the US National Cancer Institute (NCI), the National Institute of Environmental Health Sciences (NIEHS), CDC's National Institute of Occupational Safety and Health (NIOSH) and the US EPA. The AHS participant cohort includes more than 89,000 licensed commercial and private pesticide applicators and their spouses from Iowa and North Carolina. Enrolment occurred from 1993 to 1997, and data collection is ongoing. The AHS maintains on its website a list of publications associated with and using the AHS cohort (see https://aghealth.nih.gov/news/publications.html).

If the pesticide of interest has been investigated as part of the AHS (www.aghealth.org), a preliminary (Tier I/scoping) review of these studies is performed early on in the evaluation as the

[30] See https://www.epa.gov/iris/advancing-systematic-review-workshop-December-2015
[31] See http://ntp.niehs.nih.gov/pubhealth/hat/noms/index-2.html and NTP's 'Handbook for Conducting a Literature-based Assessment Using OHAT Approach for Systematic Review and Evidence Integration' at https://ntp.niehs.nih.gov/ntp/ohat/pubs/handbookjan2015_508.pdf
[32] See http://handbook.cochrane.org/
[33] See http://ehp.niehs.nih.gov/1307175/

docket (or 'dossier') is opened as part of EPA's 'Scoping' analysis. In this early Tier I/scoping phase, basic epidemiological findings and conclusions from the Agricultural Health Study are described in a Tier I/scoping document which is designed to simply summarise in brief form the pertinent conclusions of various AHS study authors if there are AHS findings relevant to a the pesticide undergoing review; this Tier I scoping review is not designed to offer detailed content, critical evaluation, or evidence synthesis, and may only touch on summarised highlights of the relevant AHS -related journal articles. If other high-quality non-AHS studies are available like those from the Children's Environmental Health and Disease Prevention Research Centres, these may be similarly summarised in this Tier I/scoping epidemiological review as well. Again, no critique or synthesis of the literature is offered. In some cases, the Tier I/scoping review may conclude that no additional epidemiological review of available evidence is further required. Alternatively, it may recommend that further review is necessary as part of a more involved Tier I/update or Tier II assessment.

A Tier I/update assessment is generally completed 1" to 3 years following the completion of the Tier I/scoping assessment and is issued, like the Tier II discussed below, along with and as part of the Draft Human Health Risk Assessment. Tier I/update assessments perform a thorough review of the available literature in the AHS. A Tier I/update assessment reviews, summarises and evaluates in a qualitative, narrative summary (including reported measures of association), the applicable studies that are listed on the AHS website.[34] Reviews are generally in the form of a narrative, focusing on the key aspects of studies and their conclusions and include EPA OPP commentary along with summary EPA OPP conclusions and recommendations for further study, if necessary.

C.2.3.2. Tier II (more extensive literature search)

A Tier II assessment is a more complete review of the available epidemiological evidence and is generally done only if the earlier Tier I/scoping document suggests a potential for a specific concern (e.g. a specific and credible exposure–disease hypothesis has been advanced and needs to be further evaluated as part of a more detailed assessment). A Tier II epidemiology assessment, similar to the Tier I/update, is generally completed 1" to 3 years following the completion of the Tier I assessment and is issued along with and as part of OPP's Draft Human Health Risk Assessment; the Tier II evaluation is considered to be a qualitative narrative review that incorporates certain elements of a systematic review. For example, a Tier II assessment will include a thorough and complete literature search that is broader than that of the Tier I/update, including not only the AHS database, but also such databases as PubMed, Web of Science, Google Scholar and Science Direct, and sometimes others using standardised, transparent and reproducible query language for which specialised professional library and information science support is obtained.[35] Evidence synthesis by EPA – albeit generally in a qualitative and narrative form – also occurs in a Tier II assessment, and overall conclusions regarding the body of epidemiological literature are made. In addition, the Tier II assessment may indicate areas in which further epidemiological data and studies with respect to specific hypothesised exposure–health outcome is of interest for future work. The Tier II assessment document will not generally attempt to integrate the epidemiological findings with other lines of evidence such as that from animal toxicology studies or information from MoAs/AOPs which may be done (separately) to some degree as part of the risk assessment. To the extent that the Tier II assessment identifies specific health outcomes putatively associated with a given pesticide, further investigation and integration across disciplines can subsequently be done as part of a more comprehensive Tier III assessment (see below).

C.2.3.3. Tier III (Full Systematic Review with Data Integration)

While a Tier II assessment examines a wide range of health outcomes appearing in the epidemiological literature that are hypothesised to be associated with a given pesticide chemical, a Tier III assessment might encompass a broader (multidisciplinary) and sometimes more quantitative/statistical evaluation of at the epidemiological evidence for the association of interest, and it attempts to more

[34] https://aghealth.nih.gov/news/publications.html

[35] Additional searches conducted under the rubric of epidemiology and biomonitoring/exposure could be done using the NHANES Exposure Reports (http://www.cdc.gov/exposurereport/); TOXNET (http://toxnet.nlm.nih.gov/); CDC NBP Biomonitoring Summaries (http://www.cdc.gov/biomonitoring/biomonitoring_summaries.html); ICICADS (http://www.inchem.org/pages/cicads.html); ATSDR Toxicological Profiles (http://www.atsdr.cdc.gov/toxprofiles/index.asp); IARC Monographs (http://monographs.iarc.fr/ENG/Monographs/PDFs/; EFSA's Draft Assessment Report Database (http://dar.efsa.europa.eu/dar-web/provision); and Biomonitoring Equivalents (https://blog.americanchemistry.com/2014/07/biomonitoring-equivalents-a-valuable-scientific-tool-for-making-better-chemical-safety-decisions/

formally integrate this with animal toxicology and MoA/AOP information. Such a Tier III assessment could take the form of a systematic review of the epidemiological literature which would be performed together with evaluation of toxicity and adverse outcome pathways. For pesticide chemicals from AHS, a Tier III analysis would also ideally incorporate the results of evaluations from other high-quality epidemiological investigations and incorporate 'Weight of the Evidence' to a greater degree to reflect a more diverse set of information sources. Results from these investigations would be used to evaluate replication and consistency with results from the AHS. Early AHS findings in a number of cases were based on only a small number of participants that had developed specific outcomes or a relatively few number of years over which the participants have been followed. As the AHS cohort ages, the release of second evaluations of some chemicals from AHS will be based on additional years of follow-up and a greater number of cases that are expected to provide a more robust basis for interpreting positive and negative associations between exposure and outcome. In addition, the AHS is increasingly generating a substantial amount of biochemical, genetic marker, and molecular data to help interpret results from the epidemiological studies. Such results may further clarify AHS findings, provide evidence for a biological basis linking exposures to outcomes, or suggest additional laboratory and observational research that might strengthen evidence for mechanisms underlying causal pathways. In addition, Tier III analyses also may take advantage of efforts to bring together information and results from international cohort studies in the International Agricultural Cohort Consortium (AgriCOH) in which AHS is a member. AgriCOH is actively working to identify opportunities and approaches for pooling data across studies, and the availability of these other cohort data should aid in assessing reproducibility and replication of exposure–outcome relationships as EPA considers, evaluates and weighs the epidemiological data.

C.2.4. OPP's open literature searching strategies and evaluation of study quality

An important aspect of the systematic review approach is the thorough, systematic, and reproducible searching of the open epidemiological literature such that much of the literature that meets the established eligibility criteria can be located.[36] OPP uses specific databases as part of their literature search and has specific guidance on their conduct (for example, OPP's open literature search guidance for human health risk assessments[37]). Evaluation of all relevant literature, application of a standardised approach for grading the strength of evidence, and clear and consistent summative language will typically be important components (NRC, 2011). In addition, a high quality exposure assessment is particularly important for environmental and occupational epidemiology studies.

A second important component of the above systematic review approach is the assessment of the validity of the findings from the identified studies. Generally speaking, the quality of epidemiological research, sufficiency of documentation of the study (study design and results), and relevance to risk assessment will be considered when evaluating epidemiology studies from the open literature for use in agency risk assessments. When considering individual study quality, various aspects of the design, conduct, analysis and interpretation of the epidemiology studies are important. These include (from US-EPA, 2016):

1) clear articulation of the hypothesis, or a clear articulation of the research objectives if the study is hypothesis-generating in nature;
2) adequate assessment of exposure for the relevant critical windows of the health effects, the range of exposure of interest for the risk assessment target population, and the availability of a dose/exposure–response trend from the study, among other qualities of exposure assessment;
3) reasonably valid and reliable outcome ascertainment (the correct identification of those with and without the health effect in the study population);
4) appropriate inclusion and exclusion criteria that result in a sample population representative of the target population, and absent systematic bias;
5) adequate measurement and analysis of potentially confounding variables, including measurement or discussion of the role of multiple pesticide exposure, or mixtures exposure in the risk estimates observed.

[36] Some advocate looking at the grey or unpublished literature to lessen potential issues associated with publication bias.
[37] See https://www.epa.gov/pesticide-science-and-assessing-pesticide-risks/guidance-identifying-selecting-and-evaluating-open and specifically p. 10 of the document 'Guidance for Considering and Using Open Literature Toxicity Studies to Support Human Health Risk Assessment' dated 28.8.2012 at https://www.epa.gov/sites/production/files/2015-07/documents/lit-studies.pdf for Special Notes on Epidemiologic Data.

6) overall characterisation of potential systematic biases in the study including errors in the selection of participation and in the collection of information, including performance of sensitivity analysis to determine the potential influence of systematic error on the risk estimates presented;
7) adequate statistical power for the exposure–outcome assessment, or evaluation of the impact of statistical power of the study if under-powered to observed effects, and appropriate discussion and/or presentation of power estimates; and
8) use of appropriate statistical modelling techniques, given the study design and the nature of the outcomes under study.

References

Ankley GT, Bennett RS, Erickson RJ, Hoff DJ, Hornung MW, Johnson RD, Mount DR, Nichols JW, Russom CL, Schmieder PK, Serrrano JA, Tietge JE and Villeneuve DL, 2010. Adverse outcome pathways: a conceptual framework to support ecotoxicology research and risk assessment. Environmental Toxicology and Chemistry, 29, 730–741.

Boobis AR, Cohen SM, Dellarco V, McGregor D, Meek ME, Vickers C, Willcocks D and Farland W, 2006. IPCS framework for analyzing the relevance of a cancer mode of action for humans. Critical Reviews in Toxicology, 36, 781–792.

Boobis AR, Doe JE, Heinrich-Hirsch B, Meek ME, Munn S, Ruchirawat M, Schlatter J, Seed J and Vickers C, 2008. IPCS framework for analyzing the relevance of a noncancer mode of action for humans. Critical Reviews in Toxicology, 38, 87–96.

Hill AB, 1965. The environment and disease: association or causation? Proceedings of the Royal Society of Medicine, 58, 295–300.

Meek, ME, Boobis A, Cote I, Dellarco V, Fotakis G, Munn S, Seed J and Vickers C, 2014. New developments in the evolution and application of the WHO/IPCS framework on mode of action/species concordance analysis. Journal of Applied Toxicology, 34, 595–606.

Meek, ME, Palermo CM, Bachman AN, North CM and Lewis RJ, 2014. Mode of action human relevance (species concordance) framework: evolution of the Bradford Hill considerations and comparative analysis of weight of evidence. Journal of Applied Toxicology, 34, 1–18.

NAS (National Academy of Sciences), 2007. Toxicity Testing on the 21st Century: A Vision and a Strategy. Board on Environmental Studies and Toxicology. Available online: https://www.nap.edu/catalog/11970/toxicity-testing-in-the-21st-century-a-vision-and-a

NAS (National Academy of Sciences), 2009. Science and decisions: advancing Risk Assessment. Board on Environmental Studies and Toxicology. Available online: http://dels.nas.edu/Report/Science-Decisions-Advancing-Risk-Assessment/12209

NAS (National Academy of Sciences), 2011. Review of the Environmental Protection Agency's Draft IRIS Assessment of Formaldehyde. Board on Environmental Studies and Toxicology. Available online: https://www.nap.edu/download/13142

Simon TW, Simons SS, Preston RJ, Boobis AR, Cohen SM, Doerrer NG, Crisp PF, McMullin TS, McQueen CA and Rowlands JC, 2014. The use of mode of action information in risk assessment: Quantitative key events/dose response framework for modelling the dose-response for key events. Critical Reviews in Toxicology, 44 (Suppl 3), 17–43.

US-EPA (Environmental Protection Agency), 2010a. Draft Framework for Incorporating Human Epidemiologic and Incident Data in Health Risk Assessment. Presented to FIFRA Scientific Advisory Panel on February 2-4 2010a. January 7. Available online: https://www.regulations.gov/document?D=EPA-HQ-OPP-2009-0851-0004

US-EPA (Environmental Protection Agency), 2010b. Transmittal of Meeting Minutes of the FIFRA Scientific Advisory Panel Meeting on the Draft Framework and Case Studies on Atrazine, Human Incidents, and the Agricultural Health Study: Incorporation of Epidemiology and Human Incident Data into Human Health Risk Assessment. MEMORANDUM dated 22 April, 2010b. SAP Minutes No. 2010-03. Available online: https://www.regulations.gov/document?D=EPA-HQ-OPP-2009-0851-0059

US-EPA (Environmental Protection Agency), 2012. Office of the Science Advisor. Risk Assessment Forum. Draft Framework for Human Health Risk Assessment to Inform Decision Making. July 12, 2012.

US-EPA (Environmental Protection Agency), 2016. Office of Pesticide Programs' Framework for Incorporating Human Epidemiologic and Incident Data in Risk Assessments for Pesticides. December 28, 2016. Available online: https://www3.epa.gov/pesticides/EPA-HQ-OPP-2008-0316-DRAFT-0075.pdf

Annex D – Effect size magnification/inflation

As described in the main text of this document, a potential source of bias may result if a study has low power. This lesser known type of bias is known 'effect size magnification'. While it is as widely known that, generally small, low-powered studies can result in false negatives since the study power is inadequate to reliably detect a meaningful effect size, it is less well known that these studies can result in inflation of effect sizes if those estimated effects are required to pass a statistical threshold (e.g. the common $p < 0.05$ threshold used for statistical significance) to be judged important, relevant, or 'discovered'. This effect – variously known as effect size magnification, the 'winners curse', truth inflation, or effect size inflation – is a phenomenon by which a 'discovered' association (i.e. one that has passed a given threshold of statistical significance to be judged meaningful) from a study with suboptimal power to make that discovery will produce an observed effect size that is artificially and systematically inflated.

Such truth inflation manifests itself as (systematic) bias away from the null in studies that achieve statistical significance in instances where studies are underpowered (Reinhart, 2015). This is because low-powered (and thus generally smaller) studies are more likely to have widely varying results and thus be more likely to be affected by random variation among individuals than larger ones. More specifically, the degree of effect size magnification that may be observed in any study depends, in part, on how widely varying the results of a study is expected to be and this depends on the power of the study; low powered studies tend to produce greater degrees of effect size magnification in results that are found to be statistically significant (or pass other threshold criteria) than higher powered studies.

As an example of this 'effect size magnification' concept and why it may come about, it is useful to imagine a trial run thousands of times with variable sample sizes. In this case, there will be a broad distribution of observed effect sizes. While the observed medians of these estimated effect sizes are expected to be close to the true effect size, the smaller trials will necessarily systematically produce a wider variation in observed effect sizes than larger trials. However, in low powered studies, only a small proportion of observed effects will pass any given (high) statistical threshold of significance and these will be only the ones with the greatest of effect sizes. Thus, when these generally smaller, low powered studies with greater random variation do indeed find a significance-triggered association as a result of passing a given statistical threshold, they are more likely to overestimate the size of that effect. What this means is that research findings of low-powered and statistically significant studies are biased in favour of finding inflated effects. As summarised by Gelman and Carlin (2014): 'when researchers use small *[underpowered]*[38] samples and noisy measurements to study small effects…, a significant result is often surprisingly likely to be in the wrong direction and to greatly overestimate an effect'. In general, it can be shown that low background (or control or natural) rates, low effect sizes of interest, and smaller sample sizes in the study end to produce lower power in the study and this leads to a greater tendency towards and magnitude of (any) inflated effect sizes.

It is important to note that the effect size inflation phenomenon is a general principle applicable to discovery science in general and is not a specific affliction or malady of epidemiology (Ioannidis, 2005; Lehrer, 2010; Button, 2013; Button et al., 2013; Gelman and Carlin, 2014; Reinhart, 2015). It is often seen in studies in pharmacology, in gene studies, in psychological studies, and in much of the most-often cited medical literature. When researchers have limited ability to increase the sample size such as in most epidemiological studies, effect size magnification is not a function or fault of the research or research design, but rather a function of how that the results of that research are interpreted by the user community. Thus, unlike other possible biases such as selection or information bias in epidemiology studies, the bias is not intrinsic to the study or its design, but rather characteristic of how that study is interpreted.

In order to determine (and quantify) the potential degree of effect size magnification for any given study that produces a statistically significant result, the reviewer must perform various power calculations. More specifically, when the association between a chemical exposure and a disease is found to be statistically significant, a power analysis can be done to determine the degree to which the statistically significant effect size estimate (e.g. odds ratio, relative risk or rate ratio) may be artificially inflated.

[38] [italics added]

In order to perform the requisite power calculation, the reviewer must know or obtain four values:

1) the number of subjects in non-exposed group;
2) the number of subjects in the exposed group;
3) the number of individuals with the disease of interest (or cases) in the non-exposed group; and
4) a target value of interest to detect a difference of a given (predetermined) size in a comparison of two groups (e.g. exposed vs. not exposed)

The first three listed values are provided in or must be obtained from the publication while the target value of interest (typically an OR or RR in epidemiology studies) is selected by the risk managers (and is ultimately a policy decision).[39] This Annex examines this effect size inflation phenomenon in a quantitative way using simulations. The annex uses two example published studies and simulations of hundreds of trials to evaluate the degree to which effect size magnification may play a role in producing biased effect sizes (such as odds ratios, rate ratios or relative risks) due to low power.

The first example uses data from Agricultural Health Study prospective cohort publication examining diazinon exposure and lung cancer and illustrates the effect size magnification issue for a calculated RR. The second example uses ever-never data from a case–control study studying malathion exposure and NHL and illustrates the effect size magnification concept from the point of view of an estimated OR.

An Example Illustrating Effect Size Magnification and Relative Risk (Jones et al. *(2015))*

The power associated with a comparison between those that are not exposed to diazinon to those that are exposed at the highest tertile (T) can be computed from the information provided in the AHS study publication 'Incidence of solid tumours among pesticide applicators exposed to the organophosphate insecticide diazinon in the Agricultural Health Study - an updated analysis' by Jones et al. (2015) for lung cancer. The number of subjects at each exposure level was provided in the article (non-exposed group: N = 17710, and T(ertile)1, T2 and T3 were categorised based on exposure distribution; specifically: N of each tertile = (2,350 + 2,770)/3 = 1,710 from the publication's Table 1 where: (a) the value of 2,350 represents the number in the lowest exposed *level* and (b) the value of 2,770 represents the number of the two highest exposed levels when the exposed subjects were dichotomously categorised. Since we have (i) the number of subjects in the reference non-exposed group = 17,710; (ii) the number of subjects in each of the exposed groups (tertiles) = 1710; and (iii) the number of diseased individuals (lung cancer) in the reference non-exposed group = 199 (from Table 3 of the cited publication), we can calculate the power of the comparisons between T1 vs non-exposed, T2 vs non-exposed and T3 vs non-exposed that were presented in the article, given the assumption that any true Rate Ratio = 1.2, 1.5, or 2.0, etc.

Here, we are interested in evaluating the power associated with the estimated background rate of 199/17710 (= 0.011237), and, as a form of sensitivity analysis, one half of this background rate (or 0.005617), and twice this rate (0.022473) for detecting (admittedly arbitrary) relative rates of (possible regulatory interest of) 1.2, 1.5, 2.0 and 3.0 among the subjects in each tertile of the diazinon exposed individuals. This analysis was performed using Stata statistical software and is shown below in both tabular and graphical format for true Rate Ratios of 1.2, 1.5, 2.0 and 3.0 for

[39] This target value is an effect size of interest, often expressed as either a relative risk (for cohort studies) or an odds rate (for case control studies). That is, the target value is generally an OR or RR of a given magnitude that the risk manager desires to detect with a given degree of confidence. The higher the OR or RR, the greater the magnitude of the estimated association between exposure and the health outcome. While there are not strict guidelines about what constitutes a 'weak' association vs a 'strong' one – and it undoubtedly can be very context-dependent – values less than or equal to about 1 (or sometimes ≤ 1.2) are considered to be 'null' or 'essentially null' (this ignores the possibility of a protective effect which in some contexts – for example, vaccination efficacy – may be appropriate to consider). Values less than 2 or 3 are often considered by some as 'weak'. Values greater than 2 (or 3) and up to about 5 might be considered 'moderate', and values greater than 5 are considered by some to be 'large'. Monson (1990) describes as a guide to the strength of association a rate ratio of 1.0–1.2 as 'None', of from 1.2 to 1.5 as 'Weak', of from 1.5 to 3.0 as 'Moderate', and of 3.0–10.0 as 'Strong'. Other authors use Cohen's criteria to describe ORs of 1.5 as 'small' and 5 as 'large', with 3.5 as 'medium' in epidemiology (Cohen and Chen, 2010). Others describe 1.5 as 'small', 2.5 as 'medium' or 'moderate', 4 as 'large' or 'strong' and 10 as 'very large' or 'very strong' (Rosenthal, 1996) Taube (1995) discusses some of the limitations of environmental epidemiology in detecting weak associations (also see invited commentary illustrating counter-arguments in Wynder (1997). It should be recognized that none of the demarcation lines are 'hard' and there can be legitimate disagreements about where these are drawn and how these are considered and interpreted. Regardless, these can be very much context-dependent and the above demarcations should not be regarded as in any way official or definitive.

1/2x-, 1x- (shown below in bold/shaded) and 2x- the (observed) background rate of 199 diseased individuals/17,710 persons[40]:

Results of power analysis for a one-sided, two-sample proportions test (α = 0.05)[(a)]

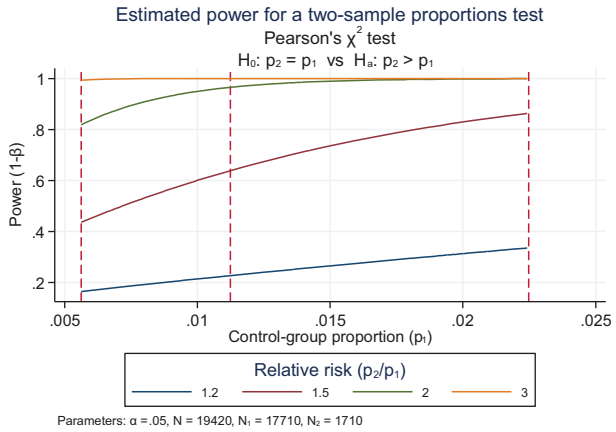
$N_{control}$	$N_{exposed}$	Proportion control[(b)]	Proportion exposed	Relative risk	Power
17,710	1,710	0.00562	0.00674	1.2	0.1634
17,710	1,710	0.00562	0.00843	1.5	0.4353
17,710	1,710	0.00562	0.01124	2.0	0.8182
17,710	1,710	0.00562	0.01685	3.0	0.9935
17,710	**1,710**	**0.01124**	**0.01348**	**1.2**	**0.2259**
17,710	**1,710**	**0.01124**	**0.01685**	**1.5**	**0.6379**
17,710	**1,710**	**0.01124**	**0.02247**	**2.0**	**0.9652**
17,710	**1,710**	**0.01124**	**0.03371**	**3.0**	**1**
17,710	1,710	0.02247	0.02697	1.2	0.3353
17,710	1,710	0.02247	0.03371	1.5	0.8632
17,710	1,710	0.02247	0.04495	2.0	0.9991
17,710	1,710	0.02247	0.06742	3.0	1

Stata code used to generate the above power calculation results: `power two proportions (‘= 0.5 * 199/17710‘= 199/17710‘= 2 * 199/17710), test(chi2) RR (1.2 1.5 2.0 3.0) n1(17710) n2(1710) one-sided table(N1:‘‘N control‘‘ N2:‘‘N exposed‘‘ p1:‘‘proportion control‘‘ p2:‘‘proportion exposed‘‘ RR:‘‘relative risk‘‘ power:‘‘power‘‘).`

(a): One-sided test α = 0.05 Ho: p2 = p1 vs Ha: p2 > p1; $N_{controls}$ = 17,710, $N_{exposed}$ = 1,710; Number of Iterations = 1,000 (data sets).

(b): Representing 1/2x-, 1x- and 2x- the observed background rate of lung cancer of 199/17710 in Jones et al. (2015). Highlighted/bolded region in table above represents power associated with this 1x observed background rate of lung cancer in cited study.

These values can be graphed as shown below[41]:



Graph showing estimated power for a (one-sided) two-sample proportions test evaluating power as a function of control-group proportion at true RRs of 1.2-, 1.5-, 2.0- and 3.0. Dashed red vertical lines represent control group proportions at 1/2x of that observed, 1x of that observed and 2x of that observed and illustrate sensitivity of the power to these background rate assumptions.

[40] The RRs of 1.2, 1.5, 2.0 and 3.0 were selected somewhat arbitrarily to illustrate the power associated with a series of relative risks that might be of interest to the risk manager/decision-maker. The values of RR or OR = 2.0 and 3.0 are considered by some to be a demarcation between weaker effect sizes and stronger effect sizes. The RR value of 1.2 is what some consider 'near to or essentially null', and the RR of 1.5 is an intermediate value between these. In determining whether the epidemiological evidence suggests a relationship between an exposure and a health outcome, a risk manager might consider the 'essentially null' RR of 1.2 from a robust study with acceptable statistical power (generally considered 80–90%) as sufficient evidence for failing to find an association and, in effect, may provide supporting evidence for a conclusion of no observable association between the exposure and the outcome.

[41] Stata code for generating the above graph: `power twoproportions (‘ = 0.5 * 199/17710‘(0.0001) ‘= 2 * 199/17710‘), test(chi2) rrisk(1.2 1.5 2.0 3.0) n1(17710) n2(1710)graph (recast(line) xline(‘= 0.5 * 199/17710‘ ‘=199/17710‘ ‘= 2 * 199/17710‘, lpattern (dash)) legend(rows(1)size(small)) ylabel(0.2(0.2)1.0)) one sided.`

As can be seen in the above table and graph, this study had a power of about 23% at 1x the background rate (control-group proportion, equal to 199 diseased individuals/17,710 subjects = 0.011237) to detect a RR of 1.2. To detect an RR of 1.5, there is about 64% power. If the true background rate were in reality twice the observed background rate (2 × 0.011237 = 0.022473), we would have about 86% power to be able to detect a RR of 1.5 and essentially 100% power to detect an RR of 2.0.[42]

Given the above, SAS was used to simulate the degree to which there may be effect size magnification (aka effect size inflation) given *true* relative risks of 1.2, 1.5, 2.0 and 3.0. The table below illustrates the power analysis for diazinon and lung cancer which shows the extent of the effect size magnification from the simulation results. The analysis presented in the table below parallels that done by Ioannidis (2008) and presented in his Table 2 for a set of hypothetical results passing the threshold of formal statistical significance to illustrate the effect size magnification concept.

SAS simulation results illustrating effect size magnification given *true* odds ratios of 1.2, 1.5, 2.0 and 3.0[(a)]

True values		N analysed data sets	Power[(b)]	Distribution of observed significant RRs			
Proportion of diseased individuals in control	RR			N	10th percentile	Median (% inflation)	90th percentile
0.005617 (1/2 × background)	1.2	1,000	0.16	157	1.6	1.7 (42)	2.0
	1.5	1,000	0.40	401	1.6	1.8 (20)	2.3
	2	1,000	0.82	823	1.7	2.1 (5)	2.8
	3	1,000	1	997	2.3	3.0 (0)	3.9
0.011237 (1 × background)	**1.2**	**1,000**	**0.22**	**224**	**1.4**	**1.6 (33)**	**1.8**
	1.5	1,000	0.63	627	1.4	1.6 (7)	2.0
	2	1,000	0.98	977	1.6	2.0 (0)	2.5
	3	1,000	1	1,000	2.5	3.0 (0)	3.6
0.022473 (2 × background)	1.2	1,000	0.33	331	1.3	1.4 (17)	1.6
	1.5	1,000	0.87	871	1.3	1.5 (0)	1.8
	2	1,000	1	1,000	1.7	2.0 (0)	2.3
	3	1,000	1	1,000	2.6	3.0 (0)	3.4

Poisson regression model was used to compare the rate of (relative risks) between the groups. The EXACT Test was used in the analysis of some data sets when the generalised Hessian matrix is not positive definite (due to a zero cases in one of the groups).
(a): One-sided test, $\alpha = 0.05$, N Controls = 17,710, N diazinon Exposed = 1,710, Number of iterations = 1,000 (data sets).
(b): The power resulting from this simulation may be close but not precisely match the power calculated from built-in procedures in statistical software such as SAS (PROC POWER) or Stata (power two-proportion). This may be due to the number of data sets simulated being of insufficient size. However, 1,000 iterations is sufficient to adequately estimate the power and to illustrate the degree of effect size magnification given a statistically significant result (here, $\alpha \leq 0.05$).

Note that – given a statistically significant result at $p < 0.05$ – the percent effect size inflation at the median of the statistically significant results varies from 0% to 42% depending on both the rate of lung cancer among individuals not exposed to diazinon (i.e. proportion of diseased individuals in the non-exposed group) and the true relative risk (ranging from 1.2 to 3.0). For example, if the **true RR** of a tertile of exposed vs non-exposed were 1.2, where the non-exposed group has a rate of lung cancer of 0.011237 (bolded row in the above table), half of the **observed** statistically significant RRs would be above the median of 1.6 and half would be below 1.6; this represents a median inflation of 33% over the true RR of 1.2 used in the simulation.

For the background rate found in the Jones et al. (2015) study (0.011237), a true RR of 1.2 that was found to be statistically significant would instead were the study to be repeated be observed to vary from 1.4 (at the 10th percentile) to 1.8 (at the 90th percentile) with the aforementioned median of 1.6. When the **true RR** is 2 or 3, the power is greater than 80% (as seen in the above table) and the median of observed RR is close to the true RR and the range of observed RRs are narrow. As the true RR increases to 3, the study's power increases such that the effect size inflation disappears and the median from the simulations indeed reflects the true RR.

[42] Said another way, if the true (but unknown) background rate were actually twice the observed background rate, we could reasonably conclude (with 86% confidence) if no statistically significant relationship was found that the true OR did not exceed 1.5.

An Example Illustrating Effect Size Magnification and Odds Ratios in an Ever/Never Analysis (Waddell, et al. 2001)

Sometimes comparisons between exposed group vs non-exposed group are presented in an 'ever/never' comparison as opposed to a comparison based on some other categorisation or grouping such as terciles or quartiles. This exposure category-based analysis might be done because there are an insufficient number of cases to break the exposure categories into small (more homogenous) exposure classifications or groupings or because the measurements of exposure are not available or are less reliable (such as in case–control studies). In these situations, we similarly need (i) the total number of subjects in non-exposed group; (ii) the number of subjects in exposed group; (iii) the number of diseased individuals in the non-exposed group in order to calculate the power of the comparison between exposed group vs non-exposed group at some; (iv) given or preselected odds ratios.

To illustrate how a power and effect size magnification analysis might be done for a case–control study using ever-never exposure categorisations, a study investigating the association between malathion and NHL (Waddell et al., 2001) was selected. Here, we have (i) the number of subjects in the reference non-exposed group = 1,018 (from Table 1: non-farmers = 243 diseased individuals + 775 non-diseased individuals); (ii) the number of subjects in the exposed group = 238 (from Table 4: malathion exposed individuals = 91 exposed cases + 147 non-exposed controls); (iii) the number of diseased individuals in the reference non-exposed group = 243 (from Table 1: 243 diseased individuals in the non-farmer or non-exposed group), we can similarly calculate the power of the comparisons between the ever vs never exposed, given the assumption that any true OR = 1.2, 1.5, 2.0, etc.

As was described above for lung cancer and diazinon, we estimated a power of 30.5% to detect an OR of 1.2 at the study-estimated NHL proportion of 0.2387 among non-farmers (non-exposed), as illustrated in the table below:

Results of power analysis for a one-sided, two-sample proportions test (α = 0.05)[(a)]

$N_{control}$	$N_{exposed}$	Proportion control[(b)]	Proportion exposed	Odds Ratio	Power
1,018	238	0.1194	0.1399	1.2	0.2279
1,018	238	0.1194	0.1689	1.5	0.647
1,018	238	0.1194	0.2133	2.0	0.9693
1,018	238	0.1194	0.2891	3.0	1
1,018	**238**	**0.2387**	**0.2734**	**1.2**	**0.3047**
1,018	**238**	**0.2387**	**0.3199**	**1.5**	**0.8149**
1,018	**238**	**0.2387**	**0.3854**	**2.0**	**0.9971**
1,018	**238**	**0.2387**	**0.4847**	**3.0**	**1**
1,018	238	0.4774	0.523	1.2	0.3522
1,018	238	0.4774	0.5781	1.5	0.8779
1,018	238	0.4774	0.6463	2.0	0.9992
1,018	238	0.4774	0.7327	3.0	1

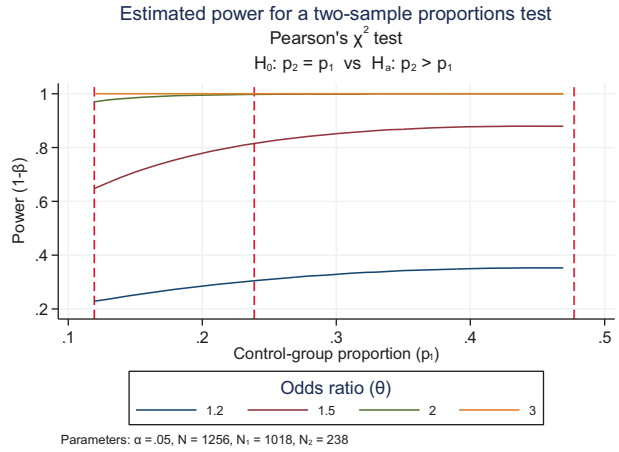
Stata code used to generate the above results: `power two-proportions ('= 0.5 * 243/1018' '= 243/1018' '= 2 * 243/1018'), test(chi2) OR (1.2 1.5 2.0 3.0) n1(1,018) n2(238) one-sided table(N1:''N control'' N2:''N exposed'' p1:''proportion control'' p2''proportion exposed'' OR:''odds ratio'' power:''power'').`

(a): One-sided test α = 0.05 Ho: p2 = p1 vs Ha: p2 > p1; $N_{controls}$ = 1,018, $N_{exposed}$ = 238, Number of iterations = 1,000 (data sets).

(b): Representing 1/2x-, 1x- and 2x- the observed background rate of lung cancer of 243/1018 in Waddell et al. (2001). Highlighted, bolded region in table above represents power associated with this 1x observed background rate of NHL in cited study.

Such power relations for malathion and NHL are graphed below[43] – as was done in the above AHS prospective cohort study for diazinon and lung cancer – with the middle vertical dotted line in the graph showing power at the NHL proportion of 0.2387 among non-farmers/non-exposed and the left-hand and right-hand vertical dashed lines representing a form of sensitivity analysis at one-half and twice the NHL proportion among non-farmers/non-exposed, respectively.

[43] Stata code for generating the graph: `power two proportions ('= 0.5 * 243/1018'(0.01) '= 2 * 243/1018'), test(chi2) OR (1.2 1.5 2.0 3.0) n1(1018) n2(238)graph(recast (line) x-line('= 0.5 * 243/1018' '= 243/1018' '= 2 * 243/1018', lpattern(dash)) legend(rows(1)size(small)) y-label(0.2(0.2)1.0)) one sided.`



Graph showing estimated power for a (one-sided) two-sample proportions test evaluating power as a function of control-group proportion at true RRs of 1.2-, 1.5-, 2.0- and 3.0. Dashed red vertical lines represent control group proportions at 1/2x of that observed, 1x of that observed and 2x of that observed and illustrates the sensitivity of the power to these background rate assumptions.

At the study-estimated NHL proportion of 0.2387 among non-farmers/non-exposed, the power (one-sided) to detect ORs of 1.2, 1.5, 2.0 and 3.0 is shown to be 30.5%, 81.5%, 99.7% and > 99.9%, respectively. Note that Waddell et al. (2001) reported an OR of 1.6 with a 95% CI of 1.2–2.2, based on 91 NHL cases who used malathion and 243 cases that were among non-farmers who did not.

Given the above, SAS was used to simulate the degree to which effect size magnification may exist given *true* odds ratios of 1.2, 1.5, 2.0 and 3.0. Below is a SAS-generated table for the power analysis for malathion and NHL showing the magnitude of the effect size magnification from the SAS-based simulation results.

SAS simulation results illustrating effect size magnification given *true* odds ratios of 1.2, 1.5, 2.0, and 3.0[(a)]

True values		N analysed data sets	Power[(b)]	Distribution of observed significant ORs			
Proportion of diseased individuals in non-exposed group	OR			N	10th percentile	Median (% inflation)	90th percentile
0.1194 (1/2 background)	1.2	1,000	0.22	220	1.4	1.5 (25)	1.8
	1.5	1,000	0.66	661	1.5	1.7 (13)	2.0
	2	1,000	0.97	972	1.6	2.0 (0)	2.5
	3	1,000	1.0	1,000	2.4	3.0 (0)	3.7
0.2387 (1× background)	**1.2**	**1,000**	**0.32**	**323**	**1.3**	**1.4 (17)**	**1.6**
	1.5	1,000	0.81	812	1.4	1.6 (7)	1.8
	2	1,000	1.0	997	1.6	2.0 (0)	2.4
	3	1,000	1.0	1,000	2.5	3.0 (0)	3.6
0.4774 (2× background)	1.2	1,000	0.34	337	1.3	1.4 (17)	1.6
	1.5	1,000	0.87	872	1.3	1.5 (0)	1.8
	2	1,000	1.0	1,000	1.6	2.0 (0)	2.5
	3	1,000	1.0	1,000	2.4	3.0 (0)	3.7

The logistic regression model was used to compute the odds ratios for the two groups. The EXACT Test was used in the analysis of some data sets when the maximum likelihood estimate did not exist (perhaps due to a zero cases in one of the groups).
(a): One-sided test, $\alpha = 0.05$, N non-exposed = 1,018, N malathion exposed = 238, N iterations = 1,000 (data sets).
(b): The power resulting from this simulation may be close but not match exactly with the power calculated from built-in procedures in statistical software such as SAS (PROC POWER) or Stata (power two-proportion). This may be due to number of data sets simulated being of insufficient size. However, 1,000 iterations are sufficient to adequately estimate the power and to illustrate the degree of effect size magnification given a statistically significant result (here, $\alpha \leq 0.05$).

Note that – given a statistically significant result at $p < 0.05$ – the median effect size varies from 1.4 to 3, depending on the NHL proportion in the non-exposed group, and the true odds ratio (ranging from 1.2 to 3.0). For example, if the true OR for a NHL proportion among non-farmers of 0.2387 was 1.2 (bolded row in the table), half of the *observed statistically significant* ORs would be above the median of 1.4 and half would be below. Further, most (90%) of the statistically significant ORs would be observed to be above 1.3, and a few (10%) would be observed even to be above 1.6.

In sum, then, the power of an epidemiological study is an important factor that should considered by regulators and others evaluating such studies. A study that is sufficiently powered will not only be more likely to detect a true effect of a given size if it is indeed present (the classic definition of power which relates to the issue of a Type II error or a false negative) but will also be less likely to magnify or exaggerate the effect if it is not there but (by chance) crosses a preselected threshold (such as the 0.05 level for statistical significance). If a study is suitably powered (say, 80% or more), the observed effect size is more likely to be a reflect a true effect size and any observed chance variation in this effect size will reflect a distribution symmetrically centred around the unknown true value. The take home message from these simulations and the original work by Ioannidis and extensions by Gelman and Carlin (2014) is that a study should be not only suitably powered to avoid a false negative (Type II error) but also suitably powered to avoid a magnification of the effect size for those effect sizes that are statistically significant (or pass some other threshold). Gelman and Carlin (2014) go further, stating that such 'retrospective design calculations may be more relevant for statistically significant findings than for nonsignificant findings. The interpretation of a statistically significant result can change drastically depending on the plausible size of the underlying effect'. Note that if a study is suitably powered, there is NO systematic risk inflation, but the effect estimates for underpowered studies that produce statistically significant effects are prone to what might be substantial risk inflation, the interpretation of which depends on realistic estimates of the true (underlying) effect.

Ideally, then, published literature studies should conduct and document power analyses. Short of that, published literature should provide adequate information for the reader to perform such power calculations (or, as Gelman and Carlin (2014) term them: (retrospective) design calculations). In the two examples provided above, the authors did provide sufficient information for the reader to calculate power and the potential for effect size magnification. This is not always the case. Sometimes information used for power calculations are only partially provided in the publications or provided information was structured in a way that does not permit such calculations.[44,45] For example, if authors use number of cases instead of level of exposure to determine tertiles or quartiles (which would be evidenced by a constant number of cases between groups) or if authors group multiple cancer outcomes together and use that number to determine tertiles, then the power (or design) calculations illustrated here are not possible since the required inputs are not able to be derived. Since the counts and data which are tabulated and reported are not necessarily standardised among authors and publications, one strong recommendation would be for publications to require reporting (even if in supplementary or online data) the necessary information to estimate power such that such evaluations can be done by both peer reviewers and interested readers.

[44] For example, in the review of the association between malathion exposure vs aggressive prostate cancer presented in the publication 'Risk of Total and Aggressive Prostate Cancer and Pesticide Use in the Agricultural Health Study' by Stella Koutros et al. (2012), the Panel was not able to calculate the power of the comparison between the malathion-exposed groups vs non-exposed group because critical information was not provided in the published article. From the publication and the supplemental document of the publication, we were able to easily find the number of *cases* in the non-exposed group (Table 2 in the main article), but the number of *subjects* in the non-exposed group or at each exposed level (i.e., quartile) appeared not to be available. We attempted to derive the number of subjects in the non-exposed group and number of subjects in each quartile from the information in Table 1 of the supplemental document of the article but were not able to do so since the information in Table 1 was presented in a way that was not consistent with many other AHS publications in that the exposed subjects were categorized into groups based on the quartiles of number of cases.

[45] Sometimes, information used for power calculations may have only been partially provided in the publications. For example, we calculated the powers associated with various thyroid cancer comparisons from the information provided in the AHS study publication 'Atrazine and Cancer Incidence Among Pesticide Applicators in the Agricultural Health Study (1994–2007)', by Laura Beane-Freeman et al. (2011). In this publication, the authors did not categorize the subjects into quartiles based on exposure but instead categorized or grouped the subjects based on the total number of *all* cancer *cases combined*. In this way, the number of cases of all types of cancer was the same between categorized groups and thus both the number of *cases* of any specific cancer of interest (e.g. thyroid, here) was not the same between groups and the number of *subjects* was not the same between groups. In this example, the publication provided (i) the reference Q1: N = 9,523, (ii) total subjects in Q2, Q3 and Q4: N = 26,834 (Table 1) and (iii) the number of thyroid cancer cases in the reference Q1 = 3 (Table 2). The exact number of subjects in each of the compared groups (Q2, Q3 or Q4) was, however, not available.

While the above analysis suggests that potential implications of the effect size inflation phenomenon are important considerations in evaluating epidemiological studies, it is important to remember a number of caveats regarding the phenomenon and how its consideration should enter into any interpretation of epidemiological studies.

- First, while this phenomenon would tend to inflate effect sizes for underpowered studies for which the effect of interest passes a statistical (or other) threshold, there are other biases that may be present that bias estimates in the other direction, *towards* the null. This bias might be referred to as effect size *suppression*. Perhaps, the most well-known of these is non-differential misclassification bias discussed in the main body of the text. This can commonly (but not always) produce predictable biases towards the null, thereby systematically under-predicting the effect size. Recognising that this is not always true and there are potentially countervailing or counteracting factors like effect size magnification (at least for small underpowered studies) is an important step forward. Specifically, underpowered studies can result in biased estimates in a direction away from the null to a degree that that can potentially offset (and possibly more than offset) any biases towards the null that may result, for example, from non-differential misclassification bias. Regardless, what is of critical importance is to recognise that adequately powered studies are necessary to be able to have at least some minimal degree of confidence in the estimate of the effect size for a statistically significant result.
- Secondly – and as stated in the main body of the text – effect size magnification is linked to a focused effort on the part of the researcher (or regulators interpreting such a study) on identifying effects that pass a given threshold of significance (e.g. $p < 0.05$) or achieve a certain size (e.g. OR > 3) when that study is underpowered. This phenomenon, then, is of most concern when a 'pre-screening' for statistical significance (or effect size). To the extent that regulators, decision-makers and others avoid acting by focusing on only those associations that 'pass' some predetermined statistical threshold and then use that effect size to evaluate and judge the magnitude of the effect without acknowledging that it might be inflated if the study is underpowered, the phenomenon is of lesser concern. Note that effect size magnification is not a function or fault of the research or research design, **but rather a function of how that research is interpreted by the user community**.
  Unfortunately, there is sometimes a tendency for attention to focus on effect sizes that are greater than a given size or that pass a certain statistical threshold and are as such 'discovered'. As recommended by Ioannidis with respect to how these 'discoveries' should be considered (Ioannidis, 2008):

'At the time of the first postulated discovery, we usually cannot tell whether an association exists at all, let alone judge its effect size. As a starting principle, one should be cautious about effect sizes. Uncertainty is not conveyed simply by CIs (no matter if these are 95%, 99% or 99.9%).

For a new proposed association, credibility and accuracy of the proposed effect varies depending on the case. One may ask the following questions: does the research community in the field adopt widely statistical significance or similar selection thresholds for claiming research findings? Did the discovery arise from a small study? Is there room for large flexibility in the analyses? Are we unprotected from selective reporting (e.g. was the protocol not fully available upfront?). Are there people or organisations interested in finding and promoting specific "positive" results? Finally, are the counteracting forces that would deflate effects minimal?'

- Thirdly, it should be remembered that the effect size inflation phenomenon is a general principle applicable to discovery science in general and is not a specific affliction or malady of epidemiology (Ioannidis, 2005; Lehrer, 2010; Button, 2013; Button et al., 2013; Reinhart, 2015). As indicated earlier, it is often seen in studies in pharmacology, in gene studies, in psychological studies, and in much of the most-often cited medical literature. Such truth inflation occurs in instances where studies are small and underpowered because such studies have widely varying results. It can be particularly problematic in instances where many researchers are performing similar studies and compete to publish 'new' or 'exciting' results (Reinhart, 2015).

Summary and Conclusions

Effect size magnification or 'truth inflation' is a phenomenon that can result in exaggerated estimates of odds ratios, relative risks or rate ratios in those instances in which these effect measures

are derived from underpowered studies in which statistical or other thresholds need to be met in order for effects to be 'discovered'. The phenomenon is not specific to epidemiology or epidemiological studies, but rather to any science in which studies tend to be small and predetermined thresholds such as those relating to effect sizes or statistical significance are used to determine whether an effect exists. As such, it is important that users of epidemiological studies recognise this issue and its potential interpretational consequences. Specifically, any discovered associations from an underpowered study that are highlighted or focused upon on the basis of passing a statistical or other similar threshold are systematically biased away from the null. While we cannot know if any specific observed effect size from a specific study is biased away from the null as a result of being a 'discovered' association that passes a statistical threshold (just as we can't say that a specific study showing non-differential misclassification will necessarily be biased towards the null), we do know that that chance favours such a bias to some degree as illustrated by the explications presented and simulations performed here. Said another way: by choosing to focus on, report, or act upon effect sizes on the basis of those effect sizes passing a statistical or other threshold, a bias is introduced since it is inevitably more likely to select those associations that are helped by chance rather than hurt by it (Yarkoni, 2009). Again, this is an issue related to how studies are interpreted by users, not one that is intrinsic to the study design nor one that is related to good scientific principles or practices.

One (partial) solution to the above issue is for the reader to cautiously interpret effect sizes in epidemiological studies that pass a prestated threshold or are statistically significant if they arise from an underpowered study, recognising that the observed effect sizes can be systematically biased away from the null. Such an approach would require that either the authors report the power of the study or that the authors provide sufficient information for the reader to do so. Effects sizes from studies with powers substantially less than 80% should be interpreted with an appropriate degree of scepticism, recognising that these may be inflated – perhaps substantially so (particularly if the power is less than 50%). The potential degree of this inflation will depend on a number of issues including background rate of the health outcome of interest, the sample size of the study and the effect size of interest. More specifically, when (a) the smaller the background rate of the health outcome of interest is low, (b) the sample size of the study is small and (c) the effect size of interest is weak, then the power of the study (to detect that effect size) will be low and the tendency towards inflated effect sizes in statistically significant results will be high. Low power studies investigating small or weak effects in populations that have a low background rate of the health outcome of interest will tend towards the greatest degree of effect size inflation. As a result, the PPR Panel recommends that epidemiological publications either incorporate such calculations or include key information such that those calculations can be performed by the reader. Specifically:

> When the association between a given pesticide exposure and a disease is found to be statistically significant, particularly in (presumed) low powered studies, data user should perform various power calculations (or a power analysis) to determine the degree to which the statistically significant effect size estimate (OR or RR) may be artificially inflated or magnified. This requires three values to be clearly reported by epidemiological studies: (i) the number of subjects in the non-exposed group (including diseased and non-diseased individuals); (ii) the number of subjects in the exposed group (including diseased and non-diseased individuals); and (iii) the number of diseased subjects in the non-exposed group. Risk managers can then select the target value of interest (typically an OR or RR) to detect a difference of a given (predetermined) effect size between the exposed and non-exposed subjects, and evaluate the degree to which effect size magnification could potentially explain the effect size that was estimated in the study of interest.

Since it appears that (i) many epidemiological studies are frequently underpowered; (ii) it is not common for authors to provide either power calculations or (sometimes) the information in publications required to do them, and (iii) the phenomenon of effect size magnification generally appears to be little recognised in the epidemiological field, the above PPR Panel recommendation will require effort on the part of researchers/grantees, publishers, and study sponsors to implement. While the above suggests that the current state of practice in this area may leave one pessimistic, an opinion piece on this topic by researcher Kate Button (Button, 2013) describing her work in Nature Reviews Neuroscience (Button et al., 2013) offered guarded reasons for optimism:

> 'Awareness of these issues is growing and acknowledging the problem is the first step to improving current practices and identifying solutions. Although issues of publication bias are difficult to solve overnight, researchers can improve the reliability of their research by adopting well-established (but

often ignored) scientific principles: Also, researchers can improve the usefulness/reliability of their research by adopting well-established (but often ignored) scientific principles:

1) Consider statistical power in the design of our studies, and in the interpretation of our results;
2) Increase the honesty with which we disclose our methods and results.
3) Make our study protocols, and analysis plans, and even our data, publically available; and
4) Work collaboratively to pool resources and increase our sample sizes and power to replicate findings.'

Although the above set of recommendations and thoughts were set in the context of sample size and neurotoxicology, they have broad applicability to any discovery science, including epidemiology. In sum, while there is much room for improvement in the conduct and reporting of epidemiological studies for them to be useful to regulatory bodies in making public health-based choices, the issues are beginning to be better defined and recognised and – going forward – there is reason for optimism.

References

Beane Freeman, LE, Rusiecki, JA, Hoppin, JA, Lubin, JH, Koutros, S, Andreotti, G, Hoar Zahm, S, Hines, CJ, Coble, JB, Barone Adesi, F, Sloan, J. Sandler, DP, Blair, A, and Alavanja, MCR. Atrazine and cancer incidence among pesticide applicators int eh agricultural health study (1994–2007). Environ Health Perspect, 119, 1253–1259.

Button K, 2013. Unreliable neuroscience? Why power matters. *The Guardian* newspaper (UK). 10 April 2013 Available online: https://www.theguardian.com/science/sifting-the-evidence/2013/apr/10/unreliable-neuroscience-power-matters [Accessed 6 September 2017]

Button K, Ioannidis JPA, Mokrysz C, Nosek BA, Flink J, Robinson ESJ and Munafo MR, 2013. Power failure: why small sample size undermines the reliability of neuroscience. Nature Reviews Neuroscience, 14, 365–376.

Cohen P and Chen S, 2010. How big is a big odds ratio: interpreting the magnitudes of odds ratios in epidemiological studies. Communications in Statistics: Simulation and Computation, 39, 860–864.

Gelman A and Carlin J, 2014. Beyond power calculations: assessing type S (sign) and type M (magnitude) errors. Perspectives on Psychological Science, 9, 641–651.

Ioannidis JP, 2005. Why most published research findings are false. PLoS Med, 2, e124.

Ioannidis JP, 2008. Why most discovered true associations are inflated. Epidemiology, 19, 640–648.

Jones RR, Barone-Adesi F, Koutros S, Lerro CC, Blair A, Lubin J, Heltshe SL, Hoppin JA, Alavanja MC and Beane Freeman LE. Incidence of solid tumours among pesticide applicators exposed to the organophosphate insecticide diazinon in the Agricultural Health Study: an updated analysis. Occupational and Environmental Medicine, 72, 496–503.

Koutros, S, Beane Freeman, LE, Lubin, JH, Heltshe, SL, Andreotti, G, Hughes-Barry, K, DelllaValle, CT, Hoppin, JA, Sandler, DP, Lynch, CF, Blair, A and Alavanja, MCR, 2013. Risk of total and aggressive prostate cancer and pesticide use in the agricultural health study. American Journal of Epidemiology, 177, 59–74.

Lehrer J, 2010. The truth wears off: is there something wrong with the scientific method. New Yorker. 13 December, 2010. Available online: http://www.newyorker.com/magazine/2010/12/13/the-truth-wears-off [Accessed September 2017]

Reinhart A, 2015. Statistics Done Wrong: the woefully complete guide. No Starch Press (San Francisco, CA).

Rosenthal JA, 1996. Qualitative descriptors of strength of association and effect size. Journal of Social Service Research, 21, 37–59.

Taubes G, 1995. Epidemiology faces its limits. Science, 269, 164–169.

Waddell BL, Zahm SH, Baris D, Weisenburger DD, Holmes F, Burmeister LF, Cantor KP and Blair A, 2001. Agricultural use of organophosphate pesticides and the risk of non-Hodgkin's lymphoma among male farmers (United States). Cancer Causes Control, 12, 509–517.

Wynder EL, 1997. Epidemiology Faces its Limits – Reply. Invited Commentary: Response to Science Article, "Epidemiology Faces Its Limits". American Journal of Epidemiology, 143, 747–749.

Yarkoni T, 2009. Ioannidis on effect size inflation, with guest appearance by Bozo the Clown. 21 November 2009. Available online: http://www.talyarkoni.org/blog/2009/11/21/ioannidis-on-effect-size-inflation-with-guest-appearance-by-bozo-the-clown/ [Accessed on 6 September 2017]

2017 年 9 月 20 日採択
doi: 10.2903/j.efsa.2017.5007

「農薬へのばく露と健康影響に関連する疫学研究の文献レビュー」の結果のフォローアップ（追跡調査）に関する PPR パネルの意見書

植物保護製剤（農薬）とその残留物に関する EFSA パネル（PPR）
Colin Ockleford, Paulien Adriaanse, Philippe Berny, Theodorus Brock, Sabine Duquesne, Sandro Grilli, Susanne Hougaard, Michael Klein, Thomas Kuhl, Ryszard Laskowski, Kyriaki Machera, Olavi Pelkonen, Silvia Pieper, Rob Smith, Michael Stemmer, Ingvar Sundh, Ivana Teodorovic, Aaldrik Tiktak, Chris J. Topping, Gerrit Wolterink, Matteo Bottai, Thorhallur Halldorsson, Paul Hamey, Marie-Odile Rambourg, Ioanna Tzoulaki, Daniele Court Marques, Federica Crivellente, Hubert Deluyker and Antonio F. Hernandez-Jerez

抄録

2013 年に EFSA は、2006 年から 2012 年までに発表された疫学研究の包括的な システマティックレビューを発表し、農薬ばく露と多くの健康影響との関連性を調査した。かなりの量の疫学的情報が得られたにもかかわらず、これらのエビデンスの多くはかなり質が低く、多くの制限が結果に影響している可能性が高いため、確固たる結論を出すことはできなかった。このように、規則（EU）No 1107/2009 に記載されている「認可基準」を満たしていない研究は、リスク評価には適していない。この科学的意見書では、植物保護製剤（農薬）とその残留物に関する EFSA パネル（PPR パネル）は、農薬疫学研究の方法論的限界を評価するよう求められており、その主な限界はばく露の特徴付けが不十分であることが原因であることが判明した。また、前向き研究ではなく症例対照研究を頻繁に使用していることも限界と考えられた。健康影響の不適切な定義や不正確さは避ける必要があり、結果の報告はいくつかのケースで改善される可能性がある。PPR パネルは、これらの限界を克服し、リスク評価への適切な利用を促進するために、農薬疫学研究の質と信頼性を向上させる方法についての勧告を提案した。パネルは、農薬の潜在的な有害性、ばく露シナリオ、ばく露評価の方法、ばく露－反応特性、リスク特性を理解するための有用な方法として、農薬観察研究のシステマティックレビューとメタアナリシスの実施（必要に応じて）を推奨した。最後に、PPR パネルは、農薬のリスク評価のために疫学的データを含む複数のエビデンスを統合し、重み付けする方法論的アプローチを提案した。生物学的妥当性は因果関係の立証に寄与することができる。

キーワード:疫学、農薬、リスク評価、品質評価、エビデンスの統合、複数のエビデンス、エビデンスの重み付け

要求者:欧州食品安全機関（European Food Safety Authority）
課題番号:EFSA-Q-2014-00481
対応:pesticides.ppr@efsa.europa.eu

謝辞:パネルは、本研究成果にサポートを提供してくれた以下の EFSA スタッフに感謝の意を表する。Andrea Terron、Andrea Altieri、Arianna Chiusolo。パネルと EFSA は、以下のヒアリング専門家の意見に謝意を表する。(1) David Miller(US-EPA)は US-EPA の経験を共有し、効果量の算出を行った。(2) 農業健康調査のための Kent Thomas (US-EPA), (3) the INSERM Report のための Marie Christine Lecomte (INSERM), Sylvaine Cordier (INSERM) and Alexis Elbaz (INSERM) , (4) エクスポゾームおよびメタボロミクスのための Toby Athersuch (Imperial College) , (5) 農薬に職業的にばく露された人間のバイオモニタリングのデータ収集のための Peter Floyd (Risk & Policy Analysts Ltd) , Ruth Bevan (IEH Consulting Ltd) ,Kate Jones (UK Health & Safety Laboratory)。最後に、EFSA は、意見書の改訂と提供された意見に対して、科学委員会と AMU ユニットに感謝する。

提案された引用:EFSA PPR パネル(植物保護製剤(農薬)とその残留物に関する EFSA パネル)、Ockleford C, Adriaanse P, Berny P, Brock T, Duquesne S, Grilli S, Hougaard S, Klein M, Kuhl T, Laskowski R, Machera K, Pelkonen O, Pieper S. Smith R, Stemmer M, Sundh I, Teodorovic I, Tiktak A, Topping CJ, Wolterink G, Bottai M, Halldorsson T, Hamey P, Rambourg M-O, Tzoulaki I, Court Marques D, Crivellente F, Deluyker H and Hernandez-Jerez AF, 2017。外部科学研究報告書「農薬へのばく露と健康影響をリンクさせた疫学研究の文献レビュー」の結論のフォローアップに関する PPR パネルの科学研究意見(EFSA Journal 2017;15(10):5007, 101 pp. https://doi.org/10.2903/j.efsa.2017.5007)」。

ISSN:1831-4732

概要

欧州食品安全機関（EFSA）は、植物保護製剤（農薬）とその残留物に関するパネル（PPR パネル）に、外部科学報告書「農薬へのばく露と健康影響を関連付ける疫学研究の文献レビュー」（Ntzani ら、2013 年）の結果のフォローアップ（追跡調査）に関する科学的意見書の作成を依頼した。この報告書は、2006 年から 2012 年の間に発表された疫学研究のシステマティックレビューとメタアナリシスに基づいており、農薬ばく露と 23 の主要なカテゴリーのヒト健康影響との間に見出された関連性をまとめたものである。最も関連性が高いのは、肝臓がん、乳がん、胃がん、筋萎縮性側索硬化症、喘息、II 型糖尿病、小児白血病、パーキンソン病であった。評価された疫学研究に内在する弱点があるため、因果関係についての結論を導き出すことはできないが、システマティックレビューでは、特定の複雑なヒトの健康に関する影響について情報を提供するための規制研究の適合性についての懸念が提起された。

PPR パネルは、農薬に関する疫学研究の質に影響を与える方法論的限界に対処するために、科学的意見書を作成した。この意見書は、規制（EC）1107/2009 の下での農薬の更新時のピアレビュープロセスを支援することのみを目的としており、あらゆる種類のヒトばく露による臨床例や中毒事故（入手可能であれば）を加えた疫学的研究の評価データが必要である。欧州における農薬へのばく露に関する疫学的データは、有効成分の最初の認可前には入手できないため、評価報告書草案（DAR）に提供することは期待できない。しかし、他の管轄区域では有効成分の使用について先行して認可されている可能性があり、その地域の疫学的データが適切だと考えられる。規則（EC）No 1107/2009 では、既存の疫学研究を含む学術的に査読された公表文献を検索することを要求している。このタイプのデータは、「更新のために提出された書類には、有効成分に関連する新しいデータと新しいリスク評価を含めるべきである」という規則（EC）1141/2010 にも準拠しており、有効成分の更新プロセスに適している。

本意見書では、疫学データをリスク評価に適切に活用するための農薬有効成分に特化した方法論的アプローチを提案し、農薬の疫学研究の質と信頼性を向上させるための提言を行った。さらに、PPR パネルは、農薬のリスク評価プロセスを改善するために、疫学的と実験毒性学のデータを統合するための方法論を議論し、提案した。

まず、本意見書では、観察による疫学研究[1]の基本的な要素を紹介し、因果関係を推論するための条件が通常満たされていることから、疫学研究において最も信頼性の高いエビデンスを提供すると考えられている介入研究との対比を行っている。主な観察による研究の計画については、農薬ばく露の詳細な記述の重要性、有効な健康影響の使用及びばく露と健康影響の関係をモデル化するための適切な統計解析の重要性が説明されている。また、外部及び内部研究の妥当性については、結果における偶然の役割を説明し、ばく露以外の要因が発見された関連性を歪めないかどうかを確認するために取り上げられてもいる。いくつかの種類のヒトのデータは、農薬のリスク評価プロセス、特にハザードの特定をサポートするのに貢献することができる。正式な疫学研究以外にも、症例集積、疾病登録、毒物管理センター情報、労働衛生監視データ、市販後の監視プログラムなどのヒトのデータの他の情報源は、特に急性の特定の健康影響の場合には、ハザードの特定に有用な情報を提供することができる。

しかし、農薬ばく露と健康影響に関する既存の疫学研究の多くは、さまざまな方法論の限界や不完全性に悩まされている（Terms of Reference（ToR）1）。パネルは、ヒトの観察環境における農薬ばく露と健康影響との関連を研究することは複雑で、疫学の他の多くの分野よりも困難であると指摘している。この複雑さは、市場に出回っている有効成分の数の多さ（欧州連合 EU で使用が認可されているものは約 480 種類）、ばく露の測定の難しさ、個々の農薬へのばく露に関する定量的及び定性的データが頻繁に欠如していることなど、農薬疫学の分野におけるいくつかの特殊な特徴に起因する。EFSA の外部科学報告書（Ntzani ら、2013 年）で実施された疫学的証拠の系統的評価では、多くの方法論的限界が指摘されている。特定の農薬に対する直接かつ詳細なばく露評価が行われていない（例えば、ジェネリック農薬の使用に対して情報不足）ため、主にばく露の特徴付けが不十分であることが、ほとんどの既存の研究の主な限界となっている。前向き研究ではなくて症例対照研究を頻繁に使用していることも限界となっている。健康影響の不十分な定義、統計解析がないこと、研究結果の質の低い報告が、いくつかの農薬疫学研究の他の限界として確認されている。これらの限界は、因果関係に関する強固な結論を導き出すことを困難にするデータの不均一性や矛盾の原因と

[1] 本意見書は、観察研究（疫学研究ともいう）と警戒データのみを扱う。これに対し、介入研究（無作為化臨床試験などの実験研究とも呼ばれる）は本意見書の対象外である。

なっている。Ntzani ら、(2013 年)が取り上げたほとんどの健康影響の効果量が小さいことを考えると、研究デザインにおけるバイアスの寄与が一役買っている可能性がある。

PPR パネルはまた、リスク評価に有益な将来の農薬疫学研究を改善するための多くの再修正(ToR 2)と勧告(ToR 3)も提供している。疫学研究の質と妥当性は、以下によって高められる。(a)ばく露の適切な評価、好ましくは個人のばく露モニタリングや特定の農薬(または農薬の組み合わせ)のバイオマーカー濃度を個人レベルで使用し、ばく露の誤分類を最小化し、用量反応評価を可能にする方法で報告すること、(b)十分に有効で信頼性の高い健康影響(アウトカム)の評価(十分に定義された臨床データまたは有効な代替物)、(c)交絡変数(健康影響(アウトカム)に影響を与える他の既知のばく露を含む)を適切に考慮すること、(d)サブグループ解析(例:性別、年齢などによる層別解析)を実施し、報告すること、環境疫学の研究のために特別に開発された多くの報告ガイドラインとチェックリストは、農薬ばく露を評価する疫学研究にとっても有用なものである。これは、修正された STROBE(STrengthening the Reporting of OBservational studies in Epidemiology)基準の拡張版が特に該当し、観察による研究の正確で完全な報告書に何を含めるべきかについての推奨事項が含まれている。

ばく露評価は、個人レベル(バイオモニタリングのような他の直接的な手段で補足することができる信頼性の高い線量計を目的とする農薬に使用することにより、特定の農薬に対する直接かつ詳細なばく露評価を行う)で改善することができる。さらに、登録されたデータを電子カルテにリンクさせることにより集団レベルでのばく露を評価することができる。これにより、これまでにないサンプルサイズの研究が可能となり、ばく露とその後の疾患に関する情報を得ることができるようになる。地理情報システム(GIS)や小規模地域調査も、住居ばく露の推定値を提供するための追加的な方法として役立つかもしれない。これらのより一般的なばく露評価は、一般的なリスク因子を特定する可能性があり、規制政策全体への情報提供と、さらなる疫学研究の対象を特定することの両方で重要になる可能性がある。オミクス技術の開発はまた、生物学的マトリックス中の外来物質や代謝物(メタボロミクス)から DNA やタンパク質との複合体(アダクトミクス)まで、幅広い分子の測定を通じてばく露評価を改善するための興味深い可能性を提示している。オミクスは、複雑な化学物質の混合物への累積ばく露に対する生物学的反応の特性やシグネチャーを測定する可能性があり、生物学的経路の理解を深めることができる。つまり、ばく露レベルに関連して、健康障害に関連する生化学的、生理学的、またはその他の変化を定量化できるバイオマーカーを使用することで、健康影響を再発見することができ、また、病気の発生のメカニズムを理解するのに役立つ。

規制リスク評価(ToR 4)に疫学的研究を組み込むことは、科学者、リスク評価者、リスク管理者にとって大きな課題である。様々な疫学研究の知見は、潜在的な健康被害と有害な健康影響との関連性を評価するために使用することができ、その結果、リスク評価のプロセスに貢献することができる。農薬ばく露とヒト健康影響との関連性に関する利用可能なデータが大量にあるが、それにもかかわらず、規制上のリスク評価へのこのような研究の影響はまだ限られている。ヒトのデータはリスク評価の多くの段階で利用できるが、同じ農薬有効成分に関する他の研究がない場合には、単一の(反複されていない)疫学研究は、質が高く、規則(EU)No 1107/2009 に記載されている「認可基準」を満たしていない限り、ハザードの特性評価に利用すべきではない。これらの「認可基準」は同規則には詳述されていないため、農薬の規制評価を支援するための疫学的研究の最適な計画と報告のために、多くの勧告が考慮されるべきである。さらなる特定のガイダンスが有用であるが、これは本意見書の ToR の範囲を超えている。システマティックレビューやメタアナリシス(必要に応じて)などのエビデンス統合技術が有用なアプローチを提供する。これらのツールは、選択基準を満たすすべての個々の研究の結果を組み合わせることで、要約データを生成し、統計検出力を高め、リスク推定の精度を向上させることができるが、個々の研究の方法論的な欠陥やバイアスを克服することはできない。観察による研究のシステマティックレビューやメタアナリシスは、これらのツールが農薬の潜在的なハザード、ばく露シナリオ、ばく露評価の方法、ばく露ー反応特性、リスク特性に関する理解を強化する情報を提供するため、リスク評価に大きな影響を与える能力を持っている。システマティックレビューもまた、毒性学的な課題に答えるための潜在的なツールと考えられているが、その方法論は、異なるエビデンスの系統に合わせて対応させる必要がある。

研究の評価はベストエビデンス統合の枠組みの中で行われるべきであり、それによって各特定の研究が持つ可能性のあるバイアスの特性と疫学的データベースの全体的な整合性の評価が示される。本意見書は、単一の疫学研究で

評価すべき研究の質のパラメータと、各パラメータの関連する程度（低、中、高）を報告している。ヒトのデータをバイアスのリスクと品質に関して整理するための第一段階として、3 つの基本的なカテゴリーが提案されている。(a) バイアスのリスクが低く、信頼性が高い／中程度、(b) バイアスのリスクが中程度で、信頼性が中程度、(c) バイアスのリスクが高く、信頼性が低いのは、結果の妥当性を低下させたり、潜在的な因果関係をほとんど解釈できないような重大な方法論的限界や欠陥があるためである。これらのカテゴリーは、EFSA の有効成分のピアレビューに基づく各エビデンスの信頼性と妥当性の評価（受容可能、補足的、許容できない）と並行して行うことを意図している。規則（EU）No 1107/2009 ヒトの健康リスクを適切に記載されている「認可基準」を満たすために、明確なデータ品質基準を満たさない疫学研究の結果に基づいてリスク評価を行うべきではない。

疫学研究は補完的なデータを提供するものであり、農薬リスク評価のために in vivo の実験動物試験、in vitro のメカニズムモデル及び in silico 技術から得られるデータと統合することができる（ToR 4）。これらすべてのエビデンスを組み合わせることで、判断の改善を目的としたヒトの健康リスクの特性評価におけるエビデンスの重み付け（WOE, Weight-of-Evidence）解析に貢献することができる。異なるデータセットは補完的であり、結論を出すことができ、その結果、1 つのエビデンス系統の別のエビデンス系統との整合性を強化するのに役立つが、それらは個別には不十分であり、ヒトの健康リスクを適切に特性評価するにあたっての課題となる可能性がある。したがって特にばく露されたヒト集団で臨床的に発現するまでに数十年かかる可能性がある農薬の慢性的な健康影響については、4 つのエビデンス（疫学、動物実験、in vitro、in silico）は強力な組み合わせとなる。

最初に検討すべき事項は、対象となる健康影響が、農薬に関する既存の毒性学的・疫学的研究でどれだけカバーされているかということである。既知の健康影響／エンドポイントについて両方のタイプの研究が利用可能な場合、リスク評価に使用する前に、両方の研究の長所と短所を評価すべきである。利用可能なヒトのエビデンス（観察疫学及び監視データ）、実験的エビデンス（動物及び in vitro のデータ）、非試験データ（in silico 研究）の信頼性が評価されたら、次のステップでは、これらのデータソースに重み付けを行う必要がある。この意見書では、リスク評価をより適切にサポートするために、すべてのエビデンスを全体的な WOE フレームワークの中で考慮する統合的なアプローチを提案している。このフレームワークは、ある系統が他の系統よりも優先されるべき時を強調するいくつかの原則に基づいている。どのデータセットを優先すべきかを決定するために、ヒトのデータと実験データの一致や不一致を評価すべきである。エビデンスの全体性を評価すべきであるが、データがヒトと実験のどちらから来たものであるかに関わらず、より信頼性の高いデータがより重要視されるべきである。より困難な状況は、研究結果が一致しない場合である。このような場合には、相違の理由を検討し、矛盾の生物学的根拠の理解をより深く理解する努力をすべきである。

農薬に関するヒトのデータは、標的臓器、用量反応関係、毒性影響の可逆性に関する完全な毒性学的データベースからの外挿に基づいて行われた推定値の妥当性を検証するのに役立ち、基準値の定義に直接的な影響を与えることなく、外挿の過程を再確認するのに役立つ。このように、農薬疫学的データは、根本的な因果関係の可能性を高めるための組織的なツールとして、改訂 Bradford Hill 基準を使用して、利用可能なデータの全体である WOE の一部を形成することができる。

目次

1. 序章

1.1. 農薬リスク評価におけるヒトの健康に関する規制データ要求

先進国の規制当局は、指定された試験プロトコールに基づいて実施される義務づけられた毒性学的研究と、ヒトへのばく露の可能性の推定値に基づいて、登録された各農薬について正式なヒトのリスク評価を行っている。

欧州連合(EU)では、植物保護剤(農薬)(PPP)の上市手続きは、欧州委員会規則第 1107/2009 号[2]で規定されている。欧州委員会規則第 283/2013 号[3]及び第 284/2013 号[4]では、有効成分及びその製剤の評価及び再評価のためのデータ要求が定められている。

有効成分の哺乳類毒性に関するデータ要求は、化学有効成分については欧州委員会規則(EU)No 283/2013 の Part A に、ウイルスを含む微生物については Part B に記載されている。農薬有効成分の要求事項に関しては、ヒトのデータ使用に関する言及は、異なるエンドポイントに関連する第 5 章の異なる章で見られる。例えば、ヒト由来物質(ミクロソームまたは無処置の細胞システム)を対象とした in vitro 代謝試験を含む毒物動態及び代謝に関するデータは、哺乳類における吸収、分布、代謝及び排泄に関する研究を扱う第 5.1 章に属し、ヒト由来物質を対象とした in vitro 遺伝毒性試験は、遺伝毒性試験に関する第 5.4 章に、ヒトのボランティアにおけるアセチルコリンエステラーゼ阻害などの特殊な研究は、神経毒性試験に関する第 5.7 章に記載されている。5.8 章では、有効成分に関する補足的な試験や、薬理学的、免疫学的な試験などのいくつかの特殊な試験について言及している。

農薬の評価プロセスは主に実験研究に基づいているが、ヒトのデータはそのプロセスに関連する情報を追加することができる。ヒトのデータに関する要求は、主に規則(EU)No 283/2013 の第 5.9 章「医療データ」にある。これには、偶発的な職業上のばく露や自傷／自殺の後の医学的報告書や、製造工場の従業員の監視などのモニタリング調査が含まれる。情報は、国の毒物管理センターからの報告書や、公表文献に掲載されている疫学的研究によって生成され、報告される。同規則は、ヒトへのばく露の影響に関する「意味のある」情報が入手可能な場合には、ばく露に関する外挿法の妥当性や、標的臓器、用量反応関係、毒性影響の可逆性に関する結論を導き出すために使用することを要求している。

規則(EU)No 1107/2009 も同様に、「入手可能で、ばく露レベルとばく露期間に関するデータが裏付けされており、公認の基準に従って実施されている場合、疫学的研究は特に価値があり、提出しなければならない」としている。しかし、承認または更新プロセス中の有効成分に特化した疫学的研究を実施する義務が申請者にはないことは明らかである。むしろ、規則(EC)No 1107/2009 によると、有効成分の承認のための書類(ドシエ)を提出する申請者は、「科学的なピアレビューを受けた公的に利用可能な文献[.....]」を提出しなければならない。これは、健康への副作用を扱った有効成分及びその関連代謝物に関するものであり、書類(ドシエ)提出日前の過去 10 年以内に発表されたものでなければならない[.....]。

特に、農薬に関する疫学的研究は、「規則(EC)No 1107/2009 の下での農薬有効成分の承認のための科学的根拠に基づいた公表文献の提出」(EFSA、2011 年 a)と題する EFSA ガイダンス「政策決定を支援するための食品・飼料安全性評価へのシステマティックレビュー方法論の適用」(EFSA、2010 年 a)の原則に沿って、文献から検索する必要がある。EFSA ガイダンスに示されているように、「有効成分、その代謝物、または植物保護製剤(農薬)のための科学的に査読された公表文献を特定し、選択するプロセス」は、アプローチが体系的な文献レビューに基づいている。

ヨーロッパにおける申請者による疫学研究やより一般的なヒトデータの提出は、特にこれまでに、不完全であったり、現行の EFSA ガイダンス(EFSA、2011 年 a)に準拠していなかったりすることがあった。これは、特定の EFSA ガイダンスに従って(疫学的)文献検索を行うことを義務付けることが比較的最近になって導入(例えば AIR-3 物質に対し)さ

[2] 植物保護製剤(農薬)の上市と理事会指令 79/117/EEC 及び 91/414/EEC の廃止に関する 2009 年 10 月 21 日の欧州議会及び理事会の規則(EC)No 1107/2009。OJ L 309, 24.11.2009, p. 1-50.

[3] 活性物質のデータ要求を定めた 2013 年 3 月 1 日の欧州委員会規則(EU)No 283/2013。
植物保護製剤(農薬)の上市に関する欧州議会の規則(EC)No 1107/2009 に基づく。OJ L 93, 3.4.2013, p. 1-84.

[4] 植物保護製剤(農薬)のデータ要求を定めた 2013 年 3 月 1 日の欧州委員会規則(EU)No 284/2013。
植物保護製剤(農薬)の上市に関する欧州議会及び理事会の規則(EC)No 1107/2009 に基づく。OJ L 93, 3.4.2013, p. 85-152.

れたことによるものであろう(規則 AIR-3:Reg.(EU)No 844/2012;ガイダンス文書 SANCO/2012/11251-rev.4)。

EU における農薬のピアレビュープロセスにおける疫学的データと毒性学的結果の統合評価は奨励されるべきであるが、まだ不足している。最近の大きな議論となった例としては、グリホサートの評価に関連したものがあり、リスク評価に疫学的研究を含めるために多大な努力がなされたが、結論としては、これらの研究はグリホサートと健康影響との間の関連性を示す非常に限定的なエビデンスを提供したに過ぎず、十分なエビデンスは得られなかった。

2,4-D のピアレビューの場合、疫学的データのほとんどはリスク評価には使用されなかった。結論として、欧州の規制システムの中では、疫学的データが農薬有効成分の承認に影響を与えた例はない。

疫学研究を含む文献検索が義務化され、ガイダンスが整備された現在(EFSA、2011 年 a)、より一貫したアプローチにより、リスク評価が容易になると考えられる。しかし、規制プロセスにおいて、このような疫学的情報をどのように評価するかについての枠組みは確立されていない。特に、これらの研究の評価に用いられる古典的な基準は、現在の規制の枠組みには含まれていない(例:研究デザイン、オッズ比と相対リスクの使用、潜在的な交絡因子、多重比較、因果関係の評価)。評価報告書草案(DAR)や更新評価報告書(RAR)の作成とピアレビューの過程で、疫学的知見を適切に使用するための特定の基準や指針が必要である。EFSA ステークホルダーワークショップ(EFSA、2015 年 a)では、ばく露に関する正確な情報を提供する上で、より強固で方法論的に健全な研究が利用可能になれば、EU における農薬規制の向上が図れると予想している。

もう一つの潜在的な課題は、有効成分の更新プロセスと疫学研究の成果との同期化である。実際、疫学研究の計画、実施、解析には多くの場合、特にデータの解釈が複雑な場合には相当な時間を必要とする。

1.2. 依頼者から提供された背景と委託条件

2013 年、欧州食品安全機関(EFSA)は、イオアニナ大学医学部が実施した外部科学報告書「農薬へのばく露と健康影響に関連する疫学的研究に関する文献レビュー」を発表した(Ntzani ら、2013 年)。この報告書は、2006 年から 2012 年の間に発表された疫学研究のシステマティックレビューに基づき、農薬ばく露と調査したあらゆる健康影響(ヒトの健康影響の 23 の主要カテゴリー)との関連性をまとめたものである。特に、農薬ばく露と以下の健康影響(肝臓がん、乳がん、胃がん、筋萎縮性側索硬化症、喘息、II 型糖尿病、小児白血病、パーキンソン病)との間の統計学的に有意な関連性が、固定効果及びランダム効果のメタアナリシスによって観察された。

膨大な数の研究論文と解析(6,000 件以上)が利用可能であるにもかかわらず、報告書の著者は、大部分の健康影響については何ら確かな結論を見出すことができなかった。この観察結果は、農薬の使用とヒト健康への悪影響の発生との関連性を評価したこれまでの研究と一致しており、そのような疫学研究は多くの限界とデータの大きな不均一性が問題となっていることを認めている。著者らは特に、疫学研究における農薬の広範な定義がメタアナリシスの結果の価値を制限していることを指摘している。また、本報告書の範囲では、農薬ばく露と特定の健康影響との間の詳細な関連付けを行うことができなかった。しかし、報告書では、農薬ばく露との関連性の可能性について、より詳細な結論を出すためにさらなる研究が必要とされる多くの健康影響を強調している。

とはいえ、外部科学報告書の結果は、ヨーロッパで発表された他の同様の研究[5,6]と一致しており、農薬ばく露とヒト健康影響との関連性について、多くの疑問や懸念を投げかけている。さらに、本報告書の結果は、疫学研究の結果をどのように農薬リスク評価に統合するかについての議論の道を開くものである。このことは、EU 規則 No 283/2013 に従って疫学的結果を評価する必要がある植物保護製剤(農薬)の承認評価を扱う EFSA のピアレビューチームにとって特に重要である。同規則では、申請者は、利用可能な場合には「意味のある」疫学的研究を提出しなければならないとされている。

この科学的意見書では、PPR パネルは、外部科学報告書(Ntzani ら、2013 年)で観察された農薬ばく露とヒト健康

[5] フランス。INSERM レポート 2013。農薬ー人に及ぼす影響

[6] 英国。COT レポート 2011 年。農薬への准職業上ばく露とがん以外の健康影響に関する疫学的文献のシステマティックレビューに関する声明及び COT 報告書 2006。農薬散布と住民や居合わせただけの者(bystanders)の健康に関する環境汚染に関する王立委員会報告書に関する共同声明。

影響との関連性と、これらの結果が規制上の農薬リスク評価の背景でどのように解釈されるかを議論する。したがって、PPR パネルは、報告書で収集された疫学研究を体系的に評価し、研究の主要なデータギャップと限界に対処し、関連する勧告を提言する。

PPR パネルは特に以下を行う。

1) 利用可能な疫学研究の質と妥当性に関して外部科学報告書で明らかにされたものに基づいて(必ずしもこれに限定されないが)、ギャップと限界のすべての情報源を収集し、レビューする。

2) 上記 1) 項で特定されたギャップと限界に基づき、調査結果の質、妥当性、信頼性を向上させ、それが農薬リスク評価にどのように影響を与えるかについて、将来の疫学調査のための潜在的な改善点を提案する。これには、研究デザイン、ばく露評価、データの質と評価、健康影響の診断分類、統計解析が含まれる。

3) 情報及び/または基準が不十分または不足している分野を特定し、リスク評価への適用を改善し最適化するために、農薬疫学的研究をどのように実施するかについての提言を行う。これらの推奨事項には、第 1) 項で明らかになったギャップと限界に基づいて、ばく露評価(バイオモニタリングデータの利用を含む)、脆弱な集団のサブグループ及び/または対象となる健康影響(生化学的、機能的、形態学的、臨床的レベルでの)の調和を含む。

4) 評価報告書草案のピアレビューの過程で、疫学的知見を実験毒性学、有害転帰経路(AOP)、作用機序などのデータと統合するとともに、WOE など、農薬のリスク評価に疫学的知見を適切に利用する方法を議論する。

PRAS ユニットは、リスク評価における疫学的研究の統合を含む EFSA の包括的な科学的分野[7]への合意に基づくアプローチについて、科学技術委員会に諮る。

1.3. 委託条件の解釈

EFSA は、検討事項(ToR)の中で、2006 年から 2012 年の間に発表された農薬へのばく露とヒトの健康影響を関連付ける疫学的研究のシステマティックレビューの結果のフォローアップについて、PPR パネルに科学的意見書を作成するよう要請している(Ntzani ら、2013 年)。EU 規則 No 283/2013 によると、疫学的データを農薬リスク評価に統合することは、EU 承認のための有効成分の DAR と RAR 及び植物保護製剤(農薬)としての使用を目的とした有効成分のピアレビュープロセスにとって重要であるとされている。

PPR パネルは、委託条件の解釈において、農薬の疫学的研究で明らかになった方法論的限界に対処し、規制上の農薬リスク評価、特に承認後の物質のリスク評価への利用を容易にするためにどのように改善するかについて、そのような研究のスポンサーに勧告を行うための科学的意見書を作成することになっている。PPR パネルは、実験的な毒性試験にもその方法論と報告の質に関連した限界があることに留意しているが、これらの限界の評価は本意見書の ToR の範囲を超えている。

この科学的意見書は、規制 1107/2009 に基づく農薬の更新時のピアレビュープロセスを支援することを目的としており、疫学的研究の評価に加えて、なんらかのヒトばく露後の臨床症例や中毒事例(入手可能な場合)がデータ要求となっている。欧州における農薬へのばく露に関する疫学的データは、有効成分の最初の承認前には入手できない(製造過程で発生した事故を除いて、その可能性は非常に低いと予想される)ため、DAR に貢献することは期待できないだろう。しかし、他の管轄で有効成分の使用について先行承認を受けている可能性があり、その分野の疫学的データが有用であると考えられる。EC 規則:(EC) No 1107/2009 では、既存の疫学的研究を検索することが期待される学術的に査読された公表文献を検索することを要求している。したがって、疫学的研究が有効成分の更新プロセスにおいてより適していることが認識されており、「更新のために提出された書類には、有効成分が指令 91/414/EEC の付録 I に最初に含まれた時から、データ要求の変更や科学的・技術的知識の変化を再確認するために、有効成分に関連する新しいデータと新しいリスク評価を含めるべきである」という EC 規則 1141/2010 の規定にも準拠している。

PPR パネルは具体的に以下のトピックに取り組む。

[7] 規則(EC) No 178/2002 の第 28 条による。

1)疫学研究の質に影響を与える固有の弱点(利用可能な農薬疫学研究のギャップと限界を含む)と、規制上の農薬リスク評価との関連性を検討する。これらの弱点にはどのように対処できるか?

2)実験動物を用いた古典的な毒性学的研究を補完する疫学的研究は、農薬リスク評価の分野でどのような貢献が期待できるか?

3)農薬有効成分に特化した方法論的アプローチとして、疫学的研究をどのように適切に活用するかについて、指摘されたギャップや限界をどのように改善するかを中心に議論し、提案する。

4)リスク評価の目的で利用可能な疫学的証拠をより良く利用するための実践への再提案と推奨を提案する。疫学的情報と実験毒性学のデータを統合するための方法論を議論し、提案する。

本意見書、特にセクション 2-4 は、科学としての疫学の基礎を論じることを意図したものではない。疫学の科学的側面を深めたいと考えている読者には、疫学の一般的な教科書(例:Rothman ら、2008 年)を読むことを勧める。

本意見書は、EU 規制の背景における農薬疫学研究にのみ焦点を当てており、一般的な科学的観点からではないことを考慮に入れるべきである。したがって、実験的な毒性試験の実際の限界と弱点については、ここでは触れていない。

1.4. 追加情報

上記のトピック 1-4(第 1.3 節)に完全に対応するために、疫学的研究の多くの関連するレビュー及び疫学の知識を持つ他の国内外の機関の経験に注意を払い、疫学を農薬のリスク評価に特に適用した。付録 A ではこれらの研究に詳細な注意を払い、この分野の理解に建設的に貢献してきた著者の経験に基づいている。また、付録 A では、いくつかの公表された研究がどれほど役に立たないかを示すために、厳密さに欠けていると批判された公表情報を記録している。このような優れた(そしてあまり良くない)実践から得られた教訓は、附属書 A を相互に参照することで本文に組み込まれている。このようにして、この科学的意見書(Scientific Opinion)は、それにもかかわらずアクセス可能なすべての裏付けとなるデータで読者を圧倒することなく、本文の議論を明確に抽出し、効果的に伝えることを目的としている。

さらに、付属書 B には、疫学研究におけるばく露評価のためのツールとしての労働安全衛生戦略におけるヒト生物学的モニタリング(HBM)の役割をさらに調査し、農薬への職業上ばく露による潜在的な健康リスクの評価に貢献するために、2015 年に EFSA が委託したプロジェクトの主な成果の要約が含まれている(Bevan ら、2017 年)。

2. 農薬に関する疫学研究の一般的枠組み

ここでは、農薬に関する疫学研究の基本的な要素を紹介し、他のタイプの研究との対比を行う。詳細については、疫学の一般的な教科書を勧める(Rothman ら、2009 年)。

2.1. 研究デザイン

疫学は、いつ、どこで、どのようにして疾患が発生したかを確認するために、ヒトまたは他の標的種の集団における健康影響の分布と決定要因を研究する。これは観察による研究や介入研究(すなわち臨床試験)[8]によって行うことができ、潜在的なリスク因子へのばく露が異なる研究グループを比較する。どちらのタイプの研究も、実験室よりも管理の行き届いていない自然環境で実施される。

[8] この見解では、「ヒトデータ」には、疫学研究とも呼ばれる観察研究が含まれ、研究者は研究参加者に影響を与えることなく、因子と健康影響との間の自然な関係を観察している。警戒データもまた、この概念に該当する。これに対して、介入研究(実験研究ともいう)は本意見の対象外であり、研究者が研究デザインの一部として介入することが大きな特徴である。

自然環境で発生した疾病の事例に関する情報は、ばく露者のみを対象とした症例報告や症例シリーズという形で体系的に記録することも可能である。症例報告や症例シリーズは、ばく露の違いによって研究グループを比較するものではないが、有用な情報、特に高濃度ばく露後の急性影響に関する情報を提供することができ、ハザードの特定に役立つ可能性がある。

無作為化臨床試験では、対象となるばく露が被験者に無作為に割り付けられ、可能な限り被験者は治療法を盲検化し、それによって特定の治療法へのばく露に関する知識に起因する潜在的なバイアスを排除する。これが介入研究と呼ばれる理由である。観察による疫学研究は臨床介入研究とは異なり、対象となるばく露が登録された被験者にランダムに割り付けられておらず、参加者はばく露について盲検化されていないことが多い。これが観察的研究と呼ばれる理由である。その結果、無作為化臨床試験は平均的な治療効果のバイアスのない推定値を提供するため、計画の点で上位にランクされている。

観察による研究におけるばく露の無作為割り付けがないことは、疾患の発生に関連する他のリスク因子がばく露者と非ばく露者の間で不均等に分布している可能性があるため、重要な課題となる。これは、既知の交絡因子を測定して説明する必要があることを意味する。しかし、未知の交絡因子は対処できないが、未知の交絡因子または測定されていない交絡因子が考慮されずに放置されている可能性が常にある。さらに、観察による研究で研究参加者が現在または過去のばく露を知らないことが多い、またはこれらを正確に記憶していないことがある（例えば、副流煙、食事摂取量、または職業上のハザード）という事実は、自己報告に基づいている場合、ばく露の推定値に偏りが生じる可能性がある。例えば、がん症例と対照者が過去に農薬にばく露されたことがあるかどうかを尋ねられたとき、過去のばく露が両群間で差がなかった場合でも、がん症例は対照者とは異なるばく露を報告する可能性は低くない。

伝統的に、観察による疫学研究計画は、生態学的研究、横断研究、症例対照研究、コホート研究のいずれかに分類される。このアプローチは、ばく露評価の質とばく露から結果への方向性を評価する能力に基づいている。これらの違いは、研究の質を大きく左右する（Rothman 及び Greenland、1998 年；Pearce、2012 年）。

- **生態学的研究**は観察研究であり、ばく露、影響（結果）、またはその両方を個人レベルではなく集団レベルで測定し、両者の相関関係を調べるものである。多くの場合、ばく露は集団レベルで測定されるが、健康登録を利用することで、個人レベルでの健康影響（がん、死亡率）を抽出することができる。これらの研究は、直接のばく露評価が困難な場合や、ばく露量の大きな対照が必要な場合（異なる国や職業間のレベルの比較）によく利用される。個人レベルでのばく露及び／または影響がないことを考えると、これらの研究は仮説を立てるのに有用であるが、一般的には、ヒトまたは実験動物を用いたより厳密な計画で結果をフォローアップ（追跡調査）する必要がある。
- **横断研究**では、ばく露と健康状態が同時に評価され、ばく露の程度が異なる群における有病率（または最近の限られた時間における罹患率）が比較される。このような研究では、現在のばく露が疾患の発症につながる関連時間枠ではないかもしれないので、ばく露と疾患の間の時間的関係は確立できない。有病率の高い症例を含めることは、（ほとんどの）横断研究の大きな欠点であり、特に慢性的な長期疾患の場合には注意が必要である。それでも、ばく露と影響が多かれ少なかれ同時に発生している場合や、ばく露が経時的に変化しない場合には、横断研究はリスク評価に有用であるかもしれない。
- **症例対照研究**では、すでに対象となる疾患（例：症例）と診断されている個人の過去のばく露の推定値と、そのような疾患のない同一集団の対照との間の関連を調べるものである。集団ベースの症例対照研究では、症例は十分に整備された集団から得られ、対照は症例が発生した時点で病気にかかっていない集団のメンバーから選ばれる。症例対照研究の利点は、前向き研究に比べてサンプル数、時間、供給源が少なくて済むことであり、ある種のがんのようなまれな疾患を研究する場合には、症例対照研究が唯一の実行可能な選択肢となることが多い。症例対照研究では、ほとんどの場合、過去のばく露は「直接的な」測定に基づいて評価されるのではなく、質問者または自己記入式のアンケートや職務記述書の肩書きや職務歴などの代用手段によって得られた想起など、より確実性の低い測定を介して評価される。症例対照研究は適切なばく露評価を可能にするかもしれないが、これらの研究はばく露を推定する際に想起バイアスに陥りやすい。その他の課題としては、適切な対照の

選択及び適切な交絡因子管理の必要性が挙げられる。

- **コホート研究**では、調査対象となる集団は、将来のある時点で特定の疾患や健康影響を発症するリスクがある個人で構成されている。ベースライン時及びその後の追跡調査（前向きコホート研究）では、関連するばく露、交絡因子及び健康影響が評価される。適切な追跡期間の後、以前に評価された対象となるリスク因子に異なるばく露を受けた人々の間で、疾患の発生頻度が比較される。したがって、コホート研究は計画としては前向きであり、対象となるリスク因子や共変量へのばく露の評価は健康影響が発生する前に測定される。したがって、コホート研究は、上記の他の計画と比較して、因果関係のより良いエビデンスを提供することができる。場合によっては、コホート研究が過去のばく露の推定値に基づいていることもある。このような回顧的ばく露評価は、直接測定に比べて精度が低く、想起バイアスがかかりやすい。その結果、コホート研究から得られるエビデンスの質は、ばく露を評価するために実際に使用された方法や共変量に関する情報が収集された詳細のレベルによって異なる。コホート研究は、比較的一般的な健康影響に関する研究に特に有用である。規模の点で十分な検出力があれば、比較的まれなばく露と健康影響に適切に対処するためにも利用できる。前向きコホート研究は、異なる臨界ばく露枠を研究するためにも不可欠である。その例として、成人になるまで一定間隔で子供を追跡する縦断的出生コホート研究がある。コホート研究では、疾患発症前の潜伏期間が長い場合、長い観察期間を必要とすることがある。このような研究は、実施するには複雑で費用がかかり、追跡調査の損失が生じやすい。

2.2. 母集団とサンプルサイズ

疫学研究の主な強みは、代理種ではなく、結論を出すべき集団の中で病気を研究することである。しかし、全集団を調査できることは稀であり、その代わりに研究の目的のために参照母集団からサンプルが引き出される。その結果、前者が後者を正確に反映していない場合、研究母集団で観察された効果量は、母集団で観察された効果量と異なる可能性がある。しかし、非代表的なサンプルで行われた観察は、そのサンプル内ではまだ有効であるかもしれないが、結果を一般集団に外挿する際には注意が必要である。

研究のために対象個人をどのように選択するかを決定した後、最小で何人の参加者を登録すべきかを決定することも必要である。研究のサンプルサイズは、十分な統計的検出力を保証するのに十分な大きさでなければならない。標準検出力（感度とも呼ばれる）は 80%で、これは、ある研究が、その効果が対象集団に実際に存在するときに、ある大きさの効果を検出する能力を意味する。言い換えれば、解析の結果から正しい結論を導き出せる確率は 80%で、それに対応する確率は 20%で、間違った結論を導き出して真の効果を見逃してしまう確率である。検出力解析は、与えられたサイズの効果を検出するのに必要な最小サンプルサイズを計算するためによく使用される。小規模サンプルは、非代表的サンプルを構成する可能性が高い。統計的検出力はリスク・インフレーション（risk inflation）と密接に関係しており、小規模または検出力不足の研究から統計的に有意な結果を解釈する際には特別な注意を払う必要がある（付属書 D を参照）。

疫学研究は、実験動物を用いた毒性学的研究と同様に、多くの場合は複数のエンドポイントを調査するように計画されているが、臨床試験は単一の仮説、例えば治療の有効性などを検証するために計画され、実施される。この点では、実験動物の毒性試験プロトコールについては、OECD の農薬に関するガイダンスでは、各投与群に登録する動物の最小数が規定されているので、同じ研究で試験される他の多数のエンドポイントのいずれに対しても、十分な検出力を保証することはできない。したがって、疫学研究と実験室研究の両方を実施する際には、研究の検出力を適切に考慮することが重要である。

2.3. ばく露

ばく露測定の質は、ばく露（用量）と特定の毒性影響との間の因果関係を正確に確認する研究の能力に影響を与える。

実験動物を用いた毒性学的試験では、用量、頻度、期間、経路などの「試験実施計画」が事前に十分に定められており、その実施状況を確認することができる。これにより、例えば 90 日間の研究では、飼料中に存在する化学物質の目

標とする(そして確認された)濃度と、試験動物が毎日摂取した飼料の量を掛け合わせることで、経口経路を介して毎日投与された外部ばく露量を把握することが可能になる。また、将来的には、重要な試験で内部ばく露量を決定しなければならない。

農薬の場合、ヒトの観察環境においてはばく露濃度、ばく露経路、ばく露期間が管理されておらず曖昧のため正確なばく露量を推定することは困難である。

意味のある関連性を調べるためには、ばく露の強度、頻度、期間を測定することが必要であることが多い。ばく露には、比較的短期間の高濃度ばく露もあれば、数週間から数年にわたる低レベルの長期ばく露もある。急性の高用量の農薬ばく露の影響は数時間から数日以内に現れるかもしれないが、慢性の低用量ばく露の影響は数年後にならないと現れないかもしれない。また、病気によっては最小限のばく露で発現することもあろうが、ばく露期間が長くなればその確率は高くなる。

異なるばく露経路(経皮、吸入、経口)では、吸収と代謝に違いが生じる可能性がある。経皮または吸入が職業上の環境でばく露される経路であることが多いが、一般集団では経口摂取(食品、水)が農薬ばく露の主要な経路である。ヒトにおける薬物動態には個人差が存在するため、吸収された外部ばく露量が類似している場合でも、異なる全身ばく露または組織/器官ばく露をもたらす可能性がある。

2.4. 健康影響

健康影響という用語は、調査中の健康に関連する疾病状態、事象、行動、または状態を指す。健康影響とは、研究の焦点となる臨床事象(通常は診断コード、すなわち国際疾病分類(ICD)10)または健康影響(すなわち死亡)として表現されるものである。健康影響データを使用する際には、十分に詳細な症例定義、症例を報告し記録するシステム、そしてこれらの事象の頻度を示す尺度が必要である。

明確な症例の定義は、どこで、いつ、誰によって診断されたかを問わず、一貫して診断されることを保証し、誤分類を避けるのに必要である。症例定義には標準的な基準が必要であり、それは臨床症状や徴候の組み合わせであり、時には感度と特異性が知られている診断検査によって補完されることもある。真の有病率または罹患率を推定するためには、検査手順全体の検出感度(すなわち、健康状態の悪い人が本当に不健康と診断される確率)を認識する必要がある。

また、臨床基準には、疾患リスクの増加と関連する他の特性(例えば、年齢、職業)も含まれている。同時に、適切に測定・定義された表現型あるいは困難な臨床結果は、調査結果の妥当性を高める。

疾患登録には、診断、治療、結果に関する患者の臨床情報が含まれている。これらの登録は定期的に患者情報を更新しているため、疫学研究に有用なデータを提供することができる。死亡率、がん、その他の全国的な健康登録は、一般的に症例定義の要件を満たしており、母集団内の偶発的な症例に関する(ほぼ)網羅的なデータを提供している。これらの健康影響は、国民健康統計データベースに記録され、分類されているものの、内容的にまだ改善の余地が多々あり、また、国ごとに異なる許容診断基準に依存していることも問題である。これは、社会的利益のために有効なデータを集積する試みを混乱させる可能性がある。登録データは有意義な解析を可能にするが、データの完全性と妥当性の程度によっては、適切な推論を行うことを困難にするかもしれない。また、データベースの存続期間中におけるコーディング規約の変更は、後ろ向きデータベース研究に影響を与える可能性がある。

疾患状態は一般的に二分変数として表現されるが、順序変数(例えば、重度、中等度、軽度、無疾患)あるいは定量的変数(例えば、標的臓器における毒性反応の分子バイオマーカーや血圧、脂質、特定タンパク質血清濃度などの生理学的測定値)として測定されることもある。

データ収集の完全性とその一貫性は、研究の信頼性に大きく寄与する。診断基準、データ保存、有用性の調和は、疫学研究の質に利益をもたらすであろう。

代替エンドポイント(surrogate endpoint)は、十分に定義された疾患エンドポイント、健康影響指標、一般的な臨床検査値(反応のバイオマーカー)等の代替として使用される。これらの指標は、臨床事象の原因経路上にあると考えられる。明白な臨床疾患とは対照的に、このような健康状態の生物学的マーカーは、微妙な不顕性の毒性力学的プロセスを検出することができるかもしれない。このような健康影響のために、詳細な定量分析プロトコールは、研究室間での

比較や再現を可能にするために特定されるべきである。AOP の使用は、症例定義における違いを強調することができる。

代替健康影響（surrogate outcomes）は付加的な情報を提供するかもしれないが、検査された代替健康影響の適合性は慎重に評価される必要がある。特に、代替健康影響の妥当性は、その主な使用制限となりうる（Ia Cour ら、2010 年）。したがって、妥当性が確認されていない代替エンドポイントの採択は避けるべきである。

健康状態が他の方法により、例えば自己記入式のアンケートや電話インタビュー、地域の記録（医療や行政のデータベース）から得られた場合、あるいは臨床検査のみで収集された場合、これらは基礎となる症例の定義を正確に反映していることを実証するために検証されるべきである。

2.5. 統計的解析と報告

疫学研究の質を保証するためには、資料、方法、結果を詳細に報告し、適切な統計解析を行うことが重要である。統計解析については、記述的統計及びばく露－健康影響の関係のモデル化に分けることができる。

2.5.1. 記述的統計

記述的統計は、ばく露尺度、健康影響、可能性のある交絡因子やその他の関連因子など、研究対象グループの重要な特徴を要約することを目的としている。記述的統計には、しばしば頻度表と調査したパラメータまたは変数の中心傾向（平均値や中央値など）及びばらつき度（分散や四分位数範囲など）の測定が含まれる。

2.5.2. ばく露－健康影響の関係のモデル化

ばく露－健康の関係のモデル化は、検討中のばく露と健康影響との間に考えられる関係を評価することを目的としている。特に、この関係によってばく露の量や様式、その他の介入因子にどのように依存しているかを評価することができる。

統計的検定は、科学的研究で発見された結果が偶然の結果として起こった可能性があるかどうかを判定する。これは、個々の所見からの結果を要約し、データのランダムエラーを考慮した後、これらの要約推定値が、例えば、ばく露群と非ばく露群の間で有意に異なるかどうかを評価することによって行われる。

二分された調査結果については、統計解析によりばく露群と対照群との間で疾患頻度に差があるかどうかを検索する。これは通常、相対的な尺度を用いて行われる。コホート研究における相対リスク（RR）は、ばく露群（または高ばく露群）と非ばく露群（または低ばく露群）を比較して、ばく露と疾患との関連性の相対的な大きさを推定する。これは、ばく露群では、非ばく露群（または低ばく露群）と比較して、病気を発症する可能性が高いことを示唆している。オッズ比（OR）は、一般的に症例対照研究や横断研究における健康影響の指標であり、症例と対照（または横断研究では罹患者と非罹患者）の間のばく露のオッズ比を表し、しばしば統計的検査で使用される相対的な尺度である。用量反応関係については、異なるレベルまたは用量のばく露を比較することによって確認できる。連続的な健康影響測定の場合、結果の平均値や中央値の変化は、分散分析やその他のパラメトリック統計を用いて、異なるばく露レベルにまたがって検討されることが多い。

統計解析は、観察された変化に統計学的有意差があるか否かを確認することであるが、いずれの結果においても慎重な再検討が必要である（Greenland ら、2016 年）。

統計的に有意差がないことの解釈。帰無仮説を棄却できなかったからといって、必ずしも関連性がないということではなく、関連性の有無はその研究の検出力が十分か否かに起因する。検出力は以下の要因に依存する。

- 標本サイズ：標本サイズが小さいと、たとえ真であっても統計的な有意差を検出するのは困難である。
- 偶然性または非ランダムな要因による個々の反応や特性のばらつき：ばらつきが大きいほど、統計的な有意性を示すのは難しい。
- 効果の大きさ、またはグループ間の観察された差の大きさ：効果の大きさが小さければ小さいほど、統計的有意性を示すのは困難である。

統計的に有意な差の解釈。統計的有意差とは、観察された差が偶然性だけによるものではないことを意味する。しかし、そのような結果はまだ慎重に検討する必要性がある。

- 生物学的関連性：帰無仮説の否定は、必ずしも関連が生物学的に意味のあるものであることを意味するわけではなく、関連が因果関係にあることを意味するわけでもない（Skelly、2011 年）。重要な問題は、観察された差の大きさ（または「効果の大きさ」）が、生物学的に関連性があると考えられるほど大きいかどうかということである。このように、統計的に有意な関連性は、生物学的に関連性があるかもしれないし、ないかもしれないし、その逆もある。統計的に有意な疫学的結果は「生物学的に関連性がない」として却下されるかもしれないが、統計的に有意でない結果が「生物学的に関連性がある」と判断されることはめったにない。研究者や規制当局は、一般的に使用されている健康影響の指標について、統計的有意性を超えて「生物学的に重要な差が最小である」というエビデンスを求めているケースが増えている。生物学的有意性の関連性を研究デザインや検出力の計算に配慮した上で、統計的有意性と同様に生物学的有意性の観点から結果を報告することは、リスク評価においてますます重要になってくるであろう（Skelly、2011 年）。これは、生物学的関連性を考慮する際に考慮すべき一般的な問題と基準を概説した EFSA Scientific Committee のガイダンス文書の対象となっている（EFSA Scientific Committee、2017 年 a）；また、エビデンスを扱うプロセスに関連した三つの主要な段階で生物学的関連性を考慮するためのフレームワークが開発されている（EFSA Scientific Committee、2017 年 b）。
- 偶発誤差（ランダムエラー）：統計的精度の評価には、研究内の偶然誤差を考慮する必要がある。偶発誤差とは、研究における予見できない部分であり、その部分が偶然性に起因する。統計的検定は、科学的研究で発見された結果が偶然性の結果として発生した確率を決定する。一般的に、研究参加者の数が増えると、中心傾向（例えば平均値）の推定値の精度（標準誤差として表現されることが多い）が上がり、研究グループ間に実際の差がある場合、統計的に有意な差を検出する能力が向上する。しかし、少なくとも理論的には、観察された結果が偶然に起因するものであり、比較されたグループ間に真の違いが存在しないという可能性が常にある（Skelly、2011 年）。多くの場合、この値は 5%（有意水準）に設定される。
- 多重検定：サンプルサイズについて議論する際に前述したように、ばく露－健康影響の関係のモデル化は、原則として仮説主導型であり、予め何を検索するかを研究目的に明記しておく必要がある。しかしながら、実際には、疫学的研究（及び実験動物を用いた毒性学的研究）では、多くの場合、同じばく露に関連して多くの異なる健康影響を調査している。多くの統計的検定が実施された場合、そのうちの 5%程度は偶然にも統計的に有意な結果が得られることがある。このような複数のエンドポイント（仮説）の検定は、偽陽性結果のリスクを高めるが、これは Bonferroni、Sidak、あるいは Benjamini-Hochberg の補正や、他の適切な方法を使用することでコントロールすることができる。しかし、これはしばしば省略される。このように、研究者が同じデータセットを対象に多くの統計的検定を行った場合、実際には何も変化はないのにも拘わらず差があるように結論づけられてしまうことがある。したがって、多くの統計結果は、さらなる検証を必要とする予備的な指標と考えることが重要である。統計的有意性と生物学的有意性に関する EFSA の見解は、統計解析から導き出される仮定は、研究デザインに関連して採択すべきであることに注意を促している（EFSA、2011 年 b）。
- 効果量の拡大：あまり知られていないとはいえ、バイアス追加の原因は、サンプルサイズが小さく、結果として統計的検出力が低いことに起因する可能性がある。このあまり知られていないバイアスの種類として知られているのは、低検出力研究から生じる「効果量の拡大」である。小規模で低検出力の研究では、研究の検出力が意味のある効果量を確実に検出するには不十分であるため、偽陰性が生じる可能性があることは一般的に知られているが、推定された効果が統計的閾値（例えば、統計的有意性の判定に使用される一般的な $p<0.05$ 閾値）を通過した場合に、これらの研究が効果量の誇張（インフレ）をもたらす可能性があることはあまり知られていない。この効果は、効果量の拡大としても知られているが、これは、「発見された」関連性（すなわち、統計的有意性のある閾値を通過したもの）を有効化することを目的とした最適下限の検出力を伴う研究から得られる現象であり、観察された効果量が、人工的かつ系統的に誇張されることを意味する。これは、小規模で検出力の低い研究は、大規模な研究よりも個人間のランダムな変動（ばらつき）の影響を受けやすいからである。数学的には、結果が統計的に有意であるという

あらかじめ決められた閾値を通過することを条件に、推定された効果量は真の効果量の偏った推定値となり、この偏りの大きさは研究の検出力に反比例している。例えば、ある試験を何千回も実施した場合、観察された効果量には広い分布があり、小規模な試験では大規模な試験よりも観察された効果量のばらつきが大きくなるが、これらの推定効果量の中央値は真の効果量に近いものになる。しかし、小規模で低検出力の研究では、観察された効果のうち、任意の（高い）統計的閾値を通過するのはごく一部であり、これらは最大の効果量を持つものだけである。したがって、ランダム変動が大きい小規模で検出力の低い研究では、与えられた統計的閾値を通過した場合、実際に有意性を重視することにより引き起こされる関連性については、その効果量を過大評価する可能性が高くなる。このことが意味するものは、小規模研究における有意な研究結果は、効果を誇張して発見することに有利になるように偏っているということである。一般的に、バックグラウンド（または対照または無処置）の割合が低く、対象となる効果量が小さく、研究の検出力が低いほど、誇張効果量の増大傾向がみられる。

しかし、この現象は、統計的有意性のための「事前スクリーニング」が行われた場合にのみ存在することに注意することが重要である。要するに、オッズ比（OR）や相対リスク（RR）のような与えられた量を推定したい場合、統計的有意性のために一連の効果量を「事前スクリーニング」すると、無効値から系統的に偏った（真の効果量よりも大きい）効果量が得られるということである。規制当局、政策決定者、その他の人々がこの方法で行動している範囲では、限りない比較と考えられるものの中から統計的に有意な結果を探し、効果の大きさを評価し判断するために統計的有意性のある閾値を超えたものを使用しているため、仮定された関連の大きさを誇張した感覚になる可能性が高い。追加の詳細といくつかの効果量のシミュレーションは、本書の付属書 D で提供されている。

交絡は、ばく露と疾病との関係が他のリスク因子の影響、すなわち交絡因子の影響にある程度起因している場合に発生する。リスク因子が実際に交絡因子として作用するためには、McNamee（2003 年）により提示（以下に図示）され、従来から認識されているいくつかの要件がある。その因子は次のようなものでなければならない。

- **被曝していない人に疾患を引き起こす原因、または原因の代替指標となること。**この条件を満たす因子は「リスク因子」と呼ばれる。
- 病気の存在とは無関係に、調査集団のばく露と正または負の相関があること。調査集団がばく露群と非ばく露群に分類されている場合、その因子が 2 つの群で異なる分布（有病率）を持っていることを意味する。
- ばく露と疾患の間の因果関係の経路に中間段階はない。

交絡は、ばく露と疾病の関係を過大または過小に評価する結果となり、2 つのリスク因子の影響が分離されていなかったり、「解放」されていなかったりするために起こる。実際には、十分に強固な場合、交絡はまた、見かけ上の関連性を逆転させることもある。例えば、農業上ばく露は多くの異なるばく露カテゴリーがあるため、農業従事者は、生物学的要因（土壌生物、家畜、農場動物）、花粉、粉塵、日光、オゾンなど、潜在的な交絡因子として作用する可能性のあるものを含む、多種多様なリスク因子に一般集団よりも多くばく露されている可能性が高い。

交絡を制御するために、研究の計画段階または解析段階の両方で、多くの手順が利用可能である。大規模な研究では、計画段階でのコントロールが好ましいことが多い。計画段階では、疫学研究者は、研究者がコントロールしたい特徴を共有する個人に研究集団を限定することができる。これは「限定」として知られており、実際には、その特性によって引き起こされる交絡の潜在的な影響を取り除くことができる。研究者が交絡をコントロールするための計画段階での 2 つ目の方法は、「マッチング」によるものである。ここでは、研究者は交絡変数に基づいて個人をマッチングさせ、交絡変数が 2 つの比較グループ間で均等に分布するようにする。

計画段階を超えて、解析段階では、層別化または統計的モデリングのいずれかの方法で交絡をコントロールすることができる。コントロールの 1 つの手段は、交絡変数（例えば、男性と女性、民族、または年齢グループ）のそれぞれの下で、関連性が別々に測定される層別化によるものである。別々の推定値は、各層で測定された推定値を重み付けすることによって、共通のオッズ比（OR）、相対リスク（RR）、または他の効果量を生成するために（必要に応じて）統計学的に「集積する」ことができる（例えば、Mantel-Haenszel アプローチを使用する）。これは、分析のサンプルサイズを小さくする代償として行うことができる。比較的簡単に実行できるが、この層別化が複数の交絡因子を同時に扱うことができないことに起因する困難が生ずるかもしれない。このような状況では、統計的なモデル化（例えば、多重ロジスティック

回帰）によってコントロールを達成することができる。

上述の研究の計画と解析の段階で交絡をコントロールするために利用可能なアプローチにかかわらず、研究者が計画で考慮しなかった変数や、データを収集しなかった変数をコントロールすることができないため、この分野で疫学研究を開始する前に、交絡因子を慎重に考慮することが重要である。

疫学研究は、公表されているかどうかにかかわらず、特定のリスク因子を誤って暗示したり、不適切に否定したりする可能性のある潜在的な交絡因子を無視しているとして、しばしば批判される。このような批判にもかかわらず、そのような可能性のある交絡因子によるバイアスの影響の大きさについての議論が提示されることはほとんどない。交絡因子は、リスク推定値に実質的な歪みを生じさせるためには、対象となるばく露に強く関連した疾患の比較的強いリスク因子でなければならないことを強調しなければならない。単に交絡の可能性を提起するだけでは十分ではなく、リスク因子がなぜ交絡因子になりやすいのか、その影響がどのようなものなのか、そしてその影響が結果の解釈にとってどれほど重要なのかを説明する説得力のある議論をしなければならない。強い相対リスクは測定されていない交絡因子によるものである可能性が低いのに対し、弱い交絡因子は、研究者が解析で測定または管理していない変数による残存交絡因子によるものである可能性があるため、相対リスク（RR）、オッズ比（OR）、リスク比、回帰係数などで測定される交絡の大きさを考慮することが重要である（US-EPA、2010 年 b）。

効果修飾。ヒトの健康に対する農薬及びその他の化学物質の影響は、すべての個人で同一であるとは考えにくい。例えば、ある特定の有効成分が成人の健康な被験者に及ぼす影響は、乳児、高齢者、妊婦に及ぼす影響と同じではない可能性がある。このように、ある化学物質にばく露された場合、ある集団の一部（サブセット）が感受性が高いことから疾患を発症する可能性が高くなる。このため、「脆弱な小集団」という用語が使用されているが、これは子供、妊婦、高齢者、重病歴のある人に加え、環境化学物質へのばく露による特別な健康リスク（薬物代謝酵素、トランスポーター、または生物学的標的の遺伝的多型のため）の対象となると同定された小集団を含む。平均効果とは、あるばく露の影響をすべての小集団で平均化したものである。しかし、様々な小集団間の関連の強さには不均一性があるかもしれない。例えば、化学物質 A へのばく露と健康影響 B との間の関連の程度は、健康な成人よりも子供の方が強く、また、ばく露時に防護服を着用している人や遺伝子型の異なる人には同様な影響は見られないかもしれない。もし不均一性が本当に存在するのであれば、全体的な関連性を示す単一の要約尺度は意味をなさず、誤解を招く可能性がある。不均質性の存在は、様々な小集団における因子と効果の間に統計的に有意な相互作用があるかどうかを検定することによって評価される。しかし、実際には、これは大きな標本サイズを必要とする。

関連因子によって定義された小集団での効果を調査することは、対象となるリスク因子のヒトの健康への影響についての知識を前進させるかもしれない。

2.6. 研究の妥当性

例えば、農薬ばく露と健康影響の間に統計的に有意な関連が観察された場合、またはそのような有意な関連が観察されなかった場合には、真の関連を歪めたり、その解釈に影響を及ぼす可能性のある要因を評価して、調査研究の妥当性も評価する必要がある。これらの不完全性は、ばく露と疾病の間の関連性を（系統的に）誤って推定することになる系統的な誤差の原因に関係している。さらに、単一の研究から得られた結果は、病気を発症するリスクのある他の集団で実施された独立した調査で再現された場合、より高い妥当性を持つことになる。

時間的シーケンス（順序）。因果関係の断定は、推定される効果に先行する原因を時間的に関与させなければならない。Rothman（2002 年）は、時間性を真に因果関係がある唯一の基準と考え、時間性の欠如は因果関係を排除する。疫学的関連の時間的順序は、時間的にはばく露が結果（効果）に先行する必要性を示唆しているが、ばく露の測定が結果の測定に先行する必要はない。この要件は、ばく露が後ろ向きに評価される場合（症例対照研究）や、結果と同時に評価される場合（横断研究）よりも、前向き研究の計画（すなわちコホート研究）では容易に満たされる。しかし、前向き研究においても、疾患の発症が遅かったり、初期の疾患形態が測定しにくかったりすると、原因と結果の時間的な順序や時間的な方向性を確認することが困難になることがある（Höfler、2005 年）。

研究の妥当性については、研究対象集団からより広い集団への結果の一般化可能性も考慮しなければならない。

前述したランダムエラーは精度の問題と考えられ、サンプリング変動の影響を受けるが、バイアスは妥当性の問題と考えられている。より具体的には、バイアスの問題は一般的に、正しい母集団パラメータが推定されているかどうかに影響を与える研究デザインまたは研究分析における方法論的な不完全性を伴う。バイアスの主なタイプには、選択バイアス、情報バイアス(想起バイアス、質問者／オブザーバーバイアスを含む)、交絡因子がある。追加の潜在的なバイアスの発生源は、すでに述べた効果量の規模である。

選択バイアスは、被験者を研究に参加させるために使用された手順や方法、被験者が研究から外れる方法や、そうでなければ研究への継続的な参加に影響を与える結果として発生する妥当性に関する系統的な誤差に関係している。

典型的には、このようなバイアスは、症例対照研究において、疾患に基づいて被験者を含める(または除外する)ことが、研究対象となる前のばく露状態と何らかの形で関連している場合に発生する。一例としては、ばく露と健康影響との間に関連性が疑われることに対する初期の広報やメディアの注目が、ばく露を受けた人はばく露を受けていない人に比べて優先的に診断される傾向があるかもしれない。選択バイアスはまた、コホート研究においても、例えば、研究から外れた人(追跡調査に参加できなくなった人、離脱した人、無回答の人)と残された人の状態が異なる場合のように、ばく露群と非ばく露群が真に比較可能でない場合に発生しうる。また、横断研究では、生存者のみを研究対象とする選択的生存により、選択バイアスが生じることがある。このようなタイプのバイアスは、一般的に研究の慎重な計画と実施によって対処できる(第 4 節、第 6 節、第 8 節も参照)。

「健康労働者効果」(HWE)は、一般的に認識されている選択バイアスであり、職域疫学研究で起こりうる特定のバイアスを示すものである:労働者は、労働力として雇用される必要があるため、一般集団からの個人よりも健康である傾向があり、そのため、一般集団から得られた集団ベースのサンプルよりも好ましい健康状態を持つことが多い。このような HWE バイアスは、観察された関連性が真の効果に比べて隠されたり、軽減されたりすることがあり、その結果、化学物質やその他の有害物質にばく露された労働者の死亡率や罹患率が低く見えることがある。

情報バイアスとは、ばく露または健康影響に関する情報が異なる研究グループから得られる方法に系統的な違いがあり、その結果、研究で測定される 1 つ以上の共変量に関して不正確な情報が得られたり、測定されたりする場合の系統的な誤差のことである。情報のバイアスは、結果として、ばく露または疾病状態のいずれかに関して誤った分類につながり、OR や RR のような疫学的効果の大きさの尺度にバイアスが生じる可能性がある。

ばく露状態の誤分類は、不正確、不十分、または不正確な測定値、被験者の不正確な自己申告、またはばく露データの不正確なコーディングに起因する可能性がある。

疾患状態の誤分類は、例えば、検査室のエラー、検出バイアス、データベース内の疾患状態の不正確な、または一貫性のないコーディング、あるいは不正確な想起から生じることがある。想起バイアスは情報バイアスの一種であり、ばく露状態(またはその逆)に応じて疾患状態の報告が異なる場合の系統的な誤りに関係している。質問者・バイアスは、質問者が個人のばく露状況を認識している場合に発生するもう一つの情報バイアスで、ばく露グループ間で意図しているかどうかに関わらず、ばく露グループ間で異なる疾患状況に関する回答を求めてしまうことがある。なぜなら、病気の被験者は、病気ではない被験者に比べて、より早い時期に発生したばく露を思い出す可能性が高いからである。これは、何らかの効果測定において帰無値(ばく露と疾病の間に関係がないという)から遠ざかるバイアスにつながる。

重要なことに、上述のような誤分類は、「差がある(differential)」場合と「差がない(non-differential)」場合があることである。これらは、(i)真にばく露されている(または病気にかかっている)人が、真にばく露されている、または病気にかかっていると正しく分類される度合いと、(ii)真にばく露されていない(または病気にかかっていない)人が、そのように正しく分類される度合いに関係している。前者は「感度」として知られているが、後者は「特異性」と呼ばれ、これらの両方がバイアスの存在と可能性のある方向性を決定する役割を果たしている。差異的誤分類とは、他の変数の値に依存する方法で誤分類が発生したことを意味し、非差異的誤分類とは、他の変数の値に依存しない誤分類を意味する。

疫学的観点から重要なことは、誤分類バイアス(差異的か非差異的か)は、そのようなばく露を分類するために使用された研究方法の感度と特異性に依存し、特定の(限定された)条件の下でバイアスの方向に予測可能な影響を及ぼすことができるということである。すなわち、研究の方法や解析の知識に基づくバイアスの方向性を特性評価する能力は、考慮される疫学的効果量(OR、RR など)が真の効果量の過小評価か過大評価かを政策決定者が判断することができ

るから、規制当局の政策決定に有用である。非差異的な誤分類バイアスが帰無値への予測可能なバイアスをもたらす（したがって、体系的に効果量を過小予測する）と一般的に想定されていますが、これは必ずしもそうではない。また、誤分類は非差異的であるという疫学研究で時々みられる一般的な仮定（これは、非差異的な誤分類バイアスが常に帰無（null）に向かっているという仮定と対になっていることもある）は、必ずしも正当化されているわけではない（例えば、Jurek ら、2005 年を参照）。

測定されていない交絡因子が結果に影響を与えると考えられる場合、研究者は感度分析を実施して、影響の範囲とその結果として生じる調整された効果測定値の範囲を推定すべきである（US-EPA、2010 年 b）。しかし、定量的感度（またはバイアス）分析は、多くの疫学研究では一般的には行われておらず、ほとんどの研究者は、論文の考察で様々な潜在的なバイアスを定性的に記述している。

疫学研究者は既知ではあるが測定されていないリスク因子によるばく露の誤分類や選択バイアスなどのバイアスの影響を推定したり、見落としや考慮されていない交絡因子が観察された効果の大きさに及ぼす影響を示すために感度分析を実施することが推奨されている（Lash ら、2009 年；Gustafson 及び McCandless、2010 年）。感度分析は、リスク評価目的で疫学データをレビューする際の基準リストに組み込まれるべきである。

3. 農薬に関する利用可能な疫学研究の主な限界事項

3.1. EFSA 外部科学研究報告書の著者が指摘した限界

EFSA 外部科学報告書（Ntzani ら、2013 年；付属書 A に要約）では、多様な健康影響を調査する疫学的研究が多数報告されている。疫学的証拠を体系的に評価する努力の中で、多くの方法論的限界が強調された。これらの限界の存在下では、確固たる結論を導き出すことはできなかったが、疫学からの裏付けとなるエビデンスが存在する結果は、今後の調査のために強調された。識別された主な限界は以下の通りである（Ntzani ら、2013 年）。

- 前向き研究の欠如と、バイアスがかかりやすい研究デザイン（症例対照研究と横断研究）の頻繁な使用。さらに、評価された研究の多くは、十分な検出力を持っていないようにみえる。
- 少なくとも疫学の他の多くの分野の疫学と比較して、詳細なばく露評価が欠如している。特定の農薬ばく露と混合ばく露に関する情報は、多くの場合、不足しており、適切なバイオマーカーはほとんど使用されていない。代わりに、多くの研究では、アンケート調査（多くの場合、妥当性が確認されていない）によって評価されたばく露の大まかな定義に頼っていた。
- 結果評価の不備（おおまかな結果の定義、自己申告性の健康影響または代替健康影響の使用）。
- 報告と解析の不備（効果推定値の解釈、交絡因子のコントロール、多重検定）。
- 選択的な報告・出版バイアスと、その他のバイアス（例：利害の対立）。

各研究結果の中で観察された結果の不均一性は、しばしば大きなものであった。しかし、不均一性は常にバイアスの結果であるとは限らず、本物である可能性があり、先験的に定義されたサブグループ解析やメタ回帰を考慮することは、エビデンス統合の努力の一環であるべきである。農薬ばく露に関して特に重要な職域研究もまた、健康労働者効果の影響を受けやすく、このバイアスによって労働者の罹患率と死亡率が一般集団よりも低くなっている。健康労働者効果は、雇用期間と追跡調査の期間が長くなるにつれて低下する傾向がある。

十分な統計力を持ち、農薬ばく露の詳細な定義、多くの健康影響に関するデータ及び明白性な報告を備えている研究は、農業健康調査（Agricultural Health Study：AHS）や他の同様の研究を除けば、稀である。これらの方法論的限界のいくつかは、農薬ばく露研究に限定されたものではなく、最も重要なことは、疫学的には特異的なものではなく、動物試験を含む他の特異的な研究分野でも観察されていることに注意することが重要である（Tsilidis ら、2013 年）。

EFSA の外部科学報告書には様々に定義された広範囲の農薬が記載されているが、この情報を研究間で調和させることは困難である。研究間での結果の不均一性は、同質性と同じくらい有益な情報になるが、情報は、反復を評価したり、要約効果量を計算したりできるように調和されている必要がある。これは、真の不均一性がある場合、異なる研究をプールすることができないことを意味するものではない。単一の研究からは限られた結論を導き出すことができるのみである。それにもかかわらず、報告書では、さらなる検討と調査に値する農薬と健康影響との間の多くの関連性が強調

されている。興味深いのは、公表されている文献のかなりの割合が EU 及びほとんどの先進国で使用が認可されていない農薬に焦点を当てているという事実である（例えば、DDT とその代謝物のみに焦点を当てた研究は、対象となる研究の 10%近くを占めている（Ntzani ら、2013 年）。これらの研究は、残留農薬として残留している可能性があることや、開発途上国で使用され続けていることから、まだ適切であるかもしれない。また、報告書では、約 5 年の期間にわたるあらゆる健康影響に関連した疫学的証拠に焦点を当てている。この報告書は、農薬－健康影響の関連性の疫学的評価の分野を記述する上で有用なものではあるが、特定の疾患と農薬の問題に完全に答えることはできない。農薬ばく露に関連する疾患エンドポイントのより詳細な解析が必要であり、そのような情報が入手できる場合には、EFSA の外部科学研究報告書でカバーされている期間よりも前に発表された研究も含めるべきである。

3.2. 研究デザインの限界

倫理的な理由から、EU では低用量の農薬ばく露の安全性を試験するための無作為化比較試験は認められていない。したがって、ヒトにおける潜在的な有害健康影響に関する情報は、観察による研究を用いて抽出する必要がある。

潜伏期間の長い疾患では、ある時点でのばく露量を測定しても、そのような疾患の発症に必要な長期ばく露量を正確に再現できない可能性がある。これは、生物学的サンプル中の濃度が一定ではなく、頻繁に変動する非持続性農薬の場合に特に重要である。したがって、尿サンプル中の単一の測定値と長期の潜伏期間の結果との間に関連性があると断定する研究は、慎重に解釈されるべきである。

Ntzani の報告書で検討された 795 件の研究のうち、38%が症例対照研究であり、32%が横断研究であった。その結果、農薬ばく露による潜在的な有害健康影響のエビデンスは、少なくとも潜伏期間が長い結果については、前向きな計画を欠いた研究に大部分が基づいている。横断研究では方向性を評価することができず、観察された関連性はしばしば逆因果関係（病気はばく露によって引き起こされたのか、それとも病気がばく露に影響を与えたのか）をもたらす可能性がある。逆因果関係は多くの疫学の分野で横断研究の潜在的な問題であるが、農薬疫学では、ほとんどの場合、病気が農薬へのばく露を引き起こすことはほとんどないため、問題にはならない。

いくつかのがんなどのまれな転帰に対しては症例対照研究が頻繁に用いられるが、その主な限界は、想起バイアスがかかりやすく、ばく露の後ろ向き評価に頼らなければならないことである。しかしながら、特にまれな転帰に対しては有用な情報を提供することはできる。症例対照研究と前向き研究の結果が一致するかどうかを調べることが重要である。例えば、トランス脂肪酸の摂取量と心血管疾患との関連を調べるために実施された研究（EFSA、2004 年）では、症例対照研究と前向き研究の両方で一貫して正の関連が報告されている。2 つの研究デザイン間の効果推定値は控えめな効果量が報告された前向き研究とは系統的に異なるが、どちらの研究デザインも同様の結論に達した。農薬に関しては、研究デザインにかかわらず、パーキンソン病と農薬ばく露との間の関連の大きさについては、同様の値が観察されている（レビューは Hernández ら、2016 年）。

3.3. 研究対象の妥当性

個人がばく露される農薬の環境的に適切な用量は、動物モデルで観察された毒性を誘発するのに必要な用量よりも低いため、関連する毒性影響は、小集団の感受性の違いとの関連で理解する必要がある。潜在的に脆弱な集団は、健康な個人よりも低用量の農薬へのばく露に対してリスクが高く、時にはばく露の敏感な時期にばく露されることもある。これは遺伝的感受性の場合であり、これはリスク評価のために説明されるべき重要な要因である（Gómez-Martín ら、2015 年）。遺伝的感受性は、毒物動態に影響を与える機能的な遺伝的多型（例えば、異物代謝酵素及び膜トランスポーターをコードする遺伝子）及び／または毒力学に影響を与える機能的な遺伝的多型（例えば、異なる受容体遺伝子多型）に大きく依存する。この遺伝的多様性は、妥当な科学的仮説に基づいて考慮されるべきである。

さまざまな障害、特に神経変性疾患（パーキンソン病、アルツハイマー病、筋萎縮性側索硬化症）は、環境因子（例えば農薬）へのばく露と結び付けられてきたが、多くの場合、病気の遺伝子構造は考慮されていない。特定の集団では、特定の遺伝子変異の有病率は 5-10%に達し、時には症例の 20%を超えることもある（Gibson ら、2017 年）ので、農薬ばく露とこれらの疾患の関連性は、研究対象となる集団内の遺伝的構造によって大きく影響を受ける可能性がある。こ

れらの疾患の多くの効果量が小さいことを考えると、研究デザインでは考慮されていない特定の遺伝子の根本的な効果が、疾患リスクの推定値を修正する可能性がある。したがって、農薬ばく露との関連は、一連の神経変性疾患に関連することが知られている一般的な遺伝的変異に照らして評価する必要があるかもしれない。しかし、遺伝的変異はそれ自体が人々を農薬ばく露の増加に向かわせるものではない。

特に注目すべきサブグループは子供であり、彼らの代謝、生理、食生活、環境化学物質へのばく露パターンは成人とは異なり、有害な影響を受けやすくなるからである。生物学的感受性の窓口はほとんどの場合不明のままであり、メカニズムによって異なると予想される。性別に基づく感受性は、農薬に関連した生殖毒性や内分泌かく乱の場合にも考慮する必要がある。これらのサブグループは現在リスクアセスメントの過程で考慮されているが、追加的な保護を提供するためには、より注意を払う必要があるかもしれない。

3.4. ばく露評価における課題

農薬に関する疫学研究の主な限界は、ばく露評価の不確実性に由来する。現在認可されているほとんどの農薬は、消失半減期が短い傾向があり、作物や季節に応じて様々な製剤を散布しなければならないという事実もその限界に含まれている。その結果、正確な評価には、これらの非難分解性化学物質の間欠的な長期ばく露を把握し、個々の農薬へのばく露を定量化する必要がある。

多くの研究では、大規模な農薬群に共通する尿中の非活性代謝物（例えば、有機リン酸塩の場合はジアルキルリン酸塩、ピレスロイドの場合は 3-フェノキシ安息香酸、ネオニコチノイドの場合は 6-クロロニコチン酸）を測定することで内部ばく露を評価している。これらのデータを、以下の理由からリスクを推測するために利用すべきではない。(a)これらの代謝物の一部は、親化合物を摂取するのではなく、食品やその他の供給源から事前に生成された代謝物を摂取することで直接ばく露を反映する可能性があり、(b)異なる親化合物農薬の効果は桁違いに異なる可能性があるからである。そのため、これらの尿中代謝物に基づく HBM データは、実際の農薬ばく露量を示す他のデータと組み合わせない限り、役に立たない可能性がある。

理想的には、内部ばく露量を示すバイオマーカーを使用して個人レベルでばく露量を定量化する必要がある。利用可能なバイオマーカーのほとんどは短期間（数時間または数日）のばく露を対象としており、長期にわたって複数のサンプルを収集するコストと困難さを考えると、多くの研究では外部ばく露量として定量化されている。外部ばく露量の定量的な推定には、ばく露の頻度と期間の両方を考慮する必要があり、グループレベルではなく個人レベルで行うことが望ましい。多くの場合、外部ばく露は以下のような代理指標を用いて定量化される。

- 一般的な農薬への潜在的ばく露または実際の使用に関連する、対象者または親族が報告した仕事、職種、作業、その他のライフスタイルの習慣。
- 特定の製品または製品群の取り扱いと、既存の農薬の記録や日誌を通じて文書化された、または栽培された作物から推定された、これらへの潜在的なばく露。
- 環境データ：環境農薬モニタリング、例えば水中、ばく露場所と考えられる特定の地理学上の地域からの距離及び／または居住期間。

多くの場合、これらの代理指標測定は、質問者によるものでも、自己申告に基づくものでもよいアンケート調査を用いて記録される。しかし、アンケートデータはしばしば個人の想起と知識に依存しており、想起バイアスや質問者や被験者によってもたらされるバイアスの影響を受ける可能性がある。これらのバイアスの原因は、アンケートがバイオマーカーに対して検証されていれば、ある程度定量化することができる（つまり、個々の質問が参加者のサンプルにおけるバイオマーカー濃度をどの程度予測しているか）。ばく露が後ろ向きに評価された場合、明らかな理由により、想起の精度が損なわれ、検証が不可能になる可能性が高くなる。ばく露が記録に基づいて評価される場合も、例えば記録が不完全であったり、不正確であったりすることで、同様の困難が生じる可能性がある。

これまでの多くの研究では、ばく露の持続時間が累積ばく露の代用として使用されることが多く、ばく露は時間的に均一で連続的であると仮定している（例えば、雇用期間）が、農薬の場合はこの仮定に疑問を呈しなければならない。一部の化学物質ではばく露パターンはかなり一定であるかもしれないが、市場に出回っている多数の農薬のばく露は、

季節や個人用保護具(PPE)、作業習慣によって異なり、多くの場合、使用の反復性は高くない。個人のレベルでは、ばく露量は日ごと、時間ごとに異なることがあり、多くの場合、複数の農薬が関与している。この時間的変動性は、生物学的半減期が短い農薬の全身ばく露において特に高い変動性をもたらし、長期にわたって個人のばく露に単一または少数の測定値を外挿する際にかなりの不確実性をもたらす可能性がある。したがって、ばく露の推定値を改善するためには、時間をかけて多くの測定を繰り返すことが必要になるかもしれない。

3.5. 不適切な、あるいは検証されていない健康影響のサロゲート

疫学研究では自己申告による健康影響が頻繁に用いられているが、その理由は、大規模なサンプルや限られた資金を用いた研究では、回答を検証することが困難であることなどが挙げられる。多くの研究では、自己報告された影響と医療記録との一致について調べられているが、このような指標の検証が不足しているために、特に大規模な集団ベースの研究では、誤分類につながる可能性があり、発見された関連性の信頼性を損なう可能性がある。

臨床的に明らかになった結果に依存すると、ばく露から疾患への毒力学的連続して進行したが、まだ明らかな臨床的疾患状態に達していない人が、疾患を持っていないと誤分類される可能性が高くなる可能性がある(Nachman ら、2011 年)。そのため、ばく露後の臨床症状の発現が遅れると、不適切な時期に臨床評価だけが用いられた場合、過小申告の原因となることがある。

発がん性の場合、無症状であるが、腫瘍性状態に進行する可能性のある前腫瘍性病変として評価されている例がある。これは、AHS における農薬ばく露と関連している有効性が未決定である単クローン性ガンマグロブリン血症(MGUS)の例であり(Landgren ら、2009 年)、この状態は悪性多発性骨髄腫に進行するリスクが年平均 1%である(Zingone 及び Kuehl、2011)。しかし、MGUS が悪性多発性骨髄腫に進行するかどうか、いつ、どのように進行するかを予測することは困難である。動物実験では、農薬ばく露が前がん病変のリスクと関連している可能性があることを示す研究があるので、前がん病変とがん病変の両方の転帰を組み合わせた疫学的解析を行うことで、そのような解析の検出力を高めることができるかもしれない。

代替健康影響は、臨床的に関連する転帰に代わる注目される選択肢である。なぜならば、同じ疾患の様々な代替健康影響が存在し、それらの健康影響はより早く発生したり、評価が容易であったりして、診断までの時間を短縮できるからである。しかし、有効な代替エンドポイントは、因果関係を予測し、対象となる健康影響を正確に予測するものでなければならない。さらに、これらの代替指標は農薬の作用機序に関連していなければならず、その予測性を裏付けるために、確立された毒物学的エンドポイントに固定されていなければならない。代替指標は健康影響と相関があるかもしれないが、健康影響に対する要因の効果を捉えていないかもしれない。これは、代替指標が臨床健康影響の因果関係または強く関連しているのではなく、付随する要因に過ぎないため、臨床健康影響の予測にはならない可能性があるからである。このように、代替健康影響の妥当性は、サロゲートを使用する上での大きな制限となりうる(Ia Cour ら、2010 年)。

しかし、有害健康影響を直接測定したのではなく、代わりに有効な代替指標を用いた疫学研究に基づいて、規制上の政策決定を下すことができるかどうかについては、懸念がある。代替エンドポイントとしての代替指標の使用は、臨床的に意味のある影響を予測する上での信頼性を立証する十分なエビデンスがある場合にのみ検討されるべきである。

3.6. 統計解析と結果の解釈

農薬と健康影響との関係に関する疫学的文献にみられる統計解析と科学的知見の解釈は、他の分野で報告された疫学研究と大きく異なるものではない。したがって、2.5 節で示した疫学研究の利点と限界は、農薬に関する疫学研究にも適用される。

農薬の疫学研究のいくつかの特徴は以下の通りである。(a)農薬へのばく露を評価する際の測定誤差の存在下での適切な統計解析を行うことが少ないこと、(b)ばく露－健康影響の関係に影響を及ぼす可能性のある他の重要な因子に関する情報が不十分なことである。これらの特徴については、次の段落で詳しく説明する。

a)測定誤差のある統計解析

ばく露量を正確に測定するための困難さは、栄養疫学や環境疫学などの疫学研究の多くの分野で頻繁にみられる。制御された実験室での実験環境の外で、短期及び長期のばく露を評価することが容易ではない。大規模集団では、個人は様々な期間、様々な強度で様々な形で様々な薬剤にばく露されている。

しかし、栄養疫学や環境疫学とは異なり、農薬疫学では、測定誤差を適切に考慮した統計解析が広く利用可能であり、このテーマに関する膨大な文献があるにもかかわらず、これまでのところほとんど利用されていない。その直接的な結果として、これらの統計的手法が利用されていたとしても、推論的結論が正確で、精密なものではなかった（Bengtson ら、2016 年；Dionisio ら、2016 年；Spiegelman、2016 年）。

b）その他の重要な関係因子に関する情報

対象となる結果に影響を与える可能性のある他の関連因子を特定して測定することは、科学のすべての分野で繰り返し起こる重要な問題である。例えば、ある薬が平均的に病気を効果的に治すということを知っていても、その薬が子供や妊婦に有害であるかもしれない。年齢、妊娠、その他の特性が薬の有効性に影響を与えるかどうかは、医師、患者、製薬会社、医薬品承認機関にとって重要な情報である。

農薬疫学は、可能性のある関連因子を慎重に特定し、正確に測定し、徹底的に評価し、ばく露と健康結果の関係におけるそれらの役割を評価する機会を提供している。最も多くの場合、関連因子は潜在的な交絡因子としてスクリーニングされている。交絡因子の影響が検出された場合には、統計解析で補正する必要があった。このことは、すでに収集されたデータや今後の研究で収集される可能性のあるデータを再検討することで、この重要な問題を明らかにするための更なる調査の余地を残している。農薬の文献における統計的手法は、主に二項確率やハザード回帰モデルなどの基本的な回帰分析の標準的な応用に限定されてきた。潜在的に有用な解析アプローチ、例えば、傾向スコアマッチング、媒介分析、因果推論などは、農薬疫学に役立つであろう（Imbens 及び Rubin、2015 年）。

4. 農薬リスク評価のための将来の疫学研究への再検討案

このセクションでは、利用可能な農薬の疫学的研究の評価方法と、規制目的に役立つようにそのような研究を改善するための提案を取り上げることを目的としている。

疫学的データの潜在的な規制上の利用の可能性を検討する際に、農薬ばく露と健康影響に関する既存の疫学的研究の多くは、個々の有効成分の評価においてその価値を制限する様々な方法論的限界や不完全性に悩まされている。農薬ばく露と健康影響に関する疫学研究は、半定量的なデータを生成するか、予測モデルのアウトプットに関して定量的なリスク評価との関連性を高めることが理想的である。これにより、疫学的な結果は、農薬のリスク評価に一般的に用いられる定量的なリスク評価に匹敵する言葉で表現できるようになる。このような疫学データを予測モデルと比較してリスク評価を行う際に、どのように考えればよいのかという疑問が生じる。現行の農薬疫学研究の結果では、正確に測定された定量的な用量反応関係はほとんど達成されていない。

農薬ばく露と健康影響に関連した疫学的証拠の質、信頼性、妥当性は、（a）個々の研究の質及び（b）利用可能なすべての研究から得られた複合的なエビデンスの評価を改善させることによって向上させることができる。

4.1 疫学研究の質の評価と報告

リスク評価に使用するために文献から疫学研究を選択する際には、疫学研究の質と関連性を考慮する必要がある。この研究の質は、次のようにして高めることができる（US-EPA、2012 年；Hernández ら、2016 年）。

a）ばく露の適切な評価、特に個人レベルでのバイオマーカー濃度が用量反応評価を可能にする方法で報告されていること。

b）合理的に有効で信頼性の高い健康影響評価（よく観察された臨床症状または有効な代替指標）。

c）交絡変数（複数の化学物質へのばく露を含む）を適切に考慮していること。

d）サブグループ解析の実施と報告（例えば、性別、年齢、民族による層別化）。

生物医学的研究が多様な限界の対象となり、その限界に苦しむことは広く受け入れられている。農薬に関する疫学

研究の計画、実施、解析における弱点の評価は、誤解を招く可能性のある結果を特定し、信頼性の高いデータを特定するために不可欠である。

ガイドラインやチェックリストは、特定の分野での最良の行動に向けて導く一連のルールや原則を提供することで、個人が特定の基準を満たすのを助けるものである。疫学的証拠の評価を助けとなるためにいくつかのツールやガイドラインが開発されているが、農薬に関する研究を評価するための特定のツールは存在しない。これらの研究には、特定の注意を必要とするばく露評価に関する特別な考慮事項があるが、既存の研究を批判的に評価するための標準的な疫学的手法を適用できる可能性がある。現行の報告ガイドラインは通常、研究にバイアスをもたらした可能性のある側面に焦点を当て、調査研究中に何が行われ、何が発見されたかを完全かつ明確に説明するために必要な最低限の情報を規定している(Simeraら、2010年)。

観察による疫学研究の品質評価のために、Newcastle-Ottawa スケール(NOS)や Research Triangle Institute (RTI)のアイテムバンクなど、多くのツールが特別に計画されている。後者は、化学物質ばく露の疫学研究のバイアスのリスクと精度を評価するための 29 の質問のチェックリストからなる実用的で検証済みのツールである。また、バイオモニタリング、環境疫学、短命化学物質(BEES-C)は、バイオモニタリングを用いて短命化学物質を評価する疫学研究の質を評価するために開発されたが(LaKind ら、2015 年)、主な要素が横断であり、より広範囲に適用可能であるため、難分解性化学物質や環境対策にも利用できる。評価スキームの開発に向けた 2 つの先行研究は、環境化学物質ばく露と神経発達に関する疫学研究に焦点を当てたものである(Amler ら、2006 年;Youngstrom ら、2011 年)。

報告の質に関しては、2008 年 6 月に発足した EQUATOR ネットワーク(Enhancing the QUAlity and Transparency Of health Research(EQUATOR)Network)は、健康調査研究の明白性と正確な報告を促進する国際的な取り組みである。現在、90 以上の報告ガイドラインがリストアップされており、その中には観察による疫学研究に特化したものもある(例:Strengthening the Reporting of OBservational studies in Epidemiology (STROBE))。STROBE ステートメントには、論文のタイトル、要約、緒言、方法、結果、考察のセクションに関連する 22 項目のチェックリストを使用して、横断、症例対照、コホート研究を含む観察による研究の正確で完全な報告に何を含めるべきかについての勧告が含まれている(von Elm ら、2007 年)。STROBE ステートメントは、著者への指示書の中で言及されている生物医学雑誌の数が増えてきており、支持されている。表 1 は、疫学研究の報告の質を評価する際に STROBE が考慮すべき主な特徴をまとめたものである。STROBE の拡張機能として、STROBE Extension to Genetic Association studies(STREGA)や分子疫学研究の評価のための STROBE-ME ステートメントがある。STROBE チェックリストではばく露と健康の影響について一般的にしか言及されていないため、PPR パネルは、農薬ばく露と健康影響の分野に特化した EQUATOR ネットワークライブラリに含めるための STROBE ステートメントの拡張版を開発することを推奨する。これは、研究者や規制機関が研究の質を批判的に評価する際に大いに役立つであろう。

表 1:疫学研究報告の質を評価するための STROBE ツールの主な特徴

STROBE ステートメントアイテム		
ファクター	**項目**	**推奨**
タイトルと概要		
	1	a) 要約のタイトルに一般的に使用される用語を用いて、研究の計画を示す。 b) 何が行われ、何が発見されたかについて、有益でバランスのとれた要約を記載する。
序章		
背景・根拠	2	報告された調査の科学的背景と根拠を説明する。
目的	3	事前に決められた仮説を含めた具体的な目的を述べる。
方法		
研究デザイン	4	研究デザインの重要な要素を論文の前の方で提示する。
設定	5	募集期間、ばく露期間、追跡調査期間、データ収集期間を含めて、環境、場所、関連する日付を記述する。
参加者	6	a) コホート研究ー参加資格の基準、参加者の選択の情報源と方法を示す。フォローアップの方法を記述する。 症例対照研究ー適格性の基準及び症例の確認及び対照の選択の情報源と方法を示す。

		症例及び対照の選択の根拠を示す。 横断研究－参加資格の基準及び参加者の選択の情報源と方法を示す。 b）コホート研究－マッチさせた研究については、マッチさせた基準とばく露者と未ばく露者の数を示す。 症例対照研究－マッチさせた研究の場合、マッチさせた基準と症例ごとの対照の数を示す。
変数	7	すべての転帰、ばく露、予測因子、潜在的交絡因子及び効果修飾因子を明確に示す。該当する場合は診断基準を示す
データソース／測定	8*	対象となる各変数について、データの出所と評価（測定）方法の詳細を述べる。複数のグループがある場合は、評価方法の類似性を記述する。
バイアス	9	バイアスの原因となる可能性のあるものに対処するための対応を記述する。
研究サイズ	10	研究規模がどのようにして決定されたか説明する。
量的変数	11	分析において量的変数がどのように扱われたかを説明する。該当する場合は、どのグループ分けが選択されたか、またその理由を説明する。
統計的手法	12	a）交絡因子をコントロールするために使用したものを含め、すべての統計的方法を記述する。 b）サブグループと相互作用を調べるために使用された方法を記述する。 c）不足しているデータにどのように対処したかを説明する。 d）コホート研究－該当する場合、追跡調査までの期間の損失がどのように対処されたかを説明する。 症例対照研究－該当する場合、症例と対照のマッチングがどのように対処されたかを説明する。 横断研究－該当する場合は、サンプリング戦略を考慮した分析方法を記述する。 e）感度分析を記述する
結果		
参加者	13*	a）研究の各段階における個人の数を報告する－例えば、適格性の可能性がある、適格性の審査を受けた、適格性があると判断された、研究に参加した、追跡調査を完了した、分析をうけた、など。 b）各段階で不参加の理由を述べる。 c）フロー図の使用を検討する。
記述データ	14*	a）研究参加者の特徴（例：人口学的、臨床的、社会的）及びばく露及び潜在的交絡因子に関する情報を提供する。 b）対象となる各変数について、データが欠落している参加者の数を示す。 c）コホート研究－追跡調査期間の要約（平均値や総量など）。
アウトカムデータ	15*	コホート研究－時間経過に伴うアウトカムの数または要約評価尺度の報告。 症例対照研究－各ばく露カテゴリーの数値、またはばく露の要約評価尺度を報告する。 横断研究－アウトカムの数または要約評価尺度の報告。
主な結果	16	a）未調整推定値及び該当する場合は交絡因子調整済み推定値とその精度を示す（例：95％信頼区間）。どの交絡因子が調整されたか及びそれらが含まれている理由を明確にする。 b）連続変数が分類された場合のカテゴリー境界を報告する。 c）関連性がある場合、相対リスクの推定値を意味のある期間の絶対リスクに変換することを検討する。
その他の分析	17	実施したその他の分析（サブグループと相互作用の分析、感度分析など）を報告する。

議論		
主要な結果	18	研究目的を参考にして、主要な結果を要約する。
限界事項	19	潜在的なバイアスや不正確さの原因を考慮に入れて、研究の限界について議論する。潜在的なバイアスの方向性と大きさについて議論する。
解釈	20	目的、限界、分析の重複、類似研究の結果、関連エビデンスなどを考慮して、結果の全体的な解釈を慎重に行う。
一般化可能性	21	研究結果の一般化可能性（外的妥当性）を考察する。
その他の情報		
資金調達	22	本研究の資金源と資金提供者の役割、また、該当する場合には、本論文の基礎となった原著研究の資金源を述べる。

*：症例対照研究では症例と対照について、また、コホート研究及び横断研究では、該当する場合には、ばく露群と非ばく露群について、別々に情報を提供する。

選択的報告は、有意性のない結果や注目されない有意な結果が公表されないことで起こる可能性がある。研究者は、有意な結果やリスクの高い推定値の選択的報告を避けるべきである。この点では、疫学研究の報告の標準化は、選択的な報告を減らすか回避するのに役立つであろう。STROBE 声明や同様の取り組みは、この目的のための有用なツールである。疫学研究の中には探索的でその場限りの性質を持つ研究もあるだろうが、これは出版物に明記されるべきであり、選択的報告は最小限に抑えられるべきであり、そうすれば疫学研究の結果は最も適切な観点から解釈される（Kavvoura ら、2007 年）。

研究の事前登録とプロトコールの事前発表は、医薬品の臨床試験における報告バイアスと出版バイアスを減らすために、いくつかの雑誌の編集者と倫理委員会によって取られている措置である。観察による疫学研究についても、できるだけ明白性をもって報告バイアスや出版バイアスを減らすことを実施するために、同様の提案がなされているが、疫学者の間ではコンセンサスが得られていない（Pearce、2011 年;Rushton、2011 年）。これとは対照的に、優れた疫学的実践を促進するための多くの取り組みが専門学会によって実施されてきた。例えば、国際疫学協会（IEA、2007 年）やオランダ疫学協会の responsible epidemiologic Research Practice（DSE、2017 年）などがその例である。

正式な疫学研究のデータの質の評価は、ばく露／影響の関連性についてのエビデンスを提供するか否かに関わらず、結果よりも個々の研究の方法論的特徴のみに基づいている。しかし、リスク評価においては、研究方法の質だけでなく、研究が提供する情報の質も評価することが重要である。実際、優れた研究であっても、正式な品質評価の際に、情報の報告が不十分なために却下されてしまうことがある。

4.2. 研究デザイン

適切なばく露評価を行い、十分に実施された前向き研究は、最も信頼性の高い情報を提供し、バイアスがかかりにくい。前向き研究が利用可能な場合には、計画があまり強固でない研究の結果が追加的な裏付けとなることがある。前向き研究がない場合は、横断研究や症例対照研究の結果を考慮すべきであるが、慎重に解釈すべきである。しかし、適切に計画された症例対照研究は、あまり適切に計画されていないコホート研究よりも優れている可能性があることは認められている。解析的アプローチは研究デザインに合致したものでなければならず、必要とされる統計的手法の仮定は慎重に評価されるべきである。

長期疾患の観察による研究は前向きであることが理想的であり、ばく露と健康影響との間の時間区分は、疾患の発症に要する時間に関して適切であるように計画されるべきである。がんや心血管疾患のように潜伏期間が長い（10 年以上）ことが多い健康影響については、健康影響の評価に先立って複数回のばく露評価を行うべきである。免疫機能障害のような潜伏期間の短い他の健康影響では、適切な時間区分は数日から数週間の範囲であり、1 回のばく露評価で十分であろう。要するに、研究の理想的な計画は、検討されている健康影響の潜伏期間に依存する。予測される潜伏期間は、追跡調査の長さとばく露量が定量化されなければならない頻度の両方を決定する。

4.3. 研究対象集団

EU の人口は 5 億人を超えており、かなり雑多であるため、低用量の農薬ばく露で影響を受ける可能性のある、より感受性の高い個人が多数含まれていることが予想される。これに対処するために、層別サンプリングでは、いくつかの主要な集団特性（性別、年齢、地理的分布、民族性、遺伝的変動など）に従って対象集団をサブグループに分割し、各サブグループ内で無作為にサンプルを採取する。これにより、調査母集団の中でサブグループをバランスよく表現することができる。

次に、脆弱な集団は、小集団解析または感度分析のいずれかを用いて疫学研究で調査されるべきである。しかし、そのような解析は事前に行う必要がある。事後的なサブグループ感度分析の場合、統計的閾値はそれに応じて調整されるべきで、結果の再現性はそれに従うべきである。脆弱な小集団のエビデンスは、理想的には、ばく露のバイオマーカー、前駆症状、経時的な疾患発生率の評価を含む前向き研究を必要とする。

多くの人が前臨床状態にあり、用量反応曲線の下端に敏感に反応してしまうため、集団における毒性による疾病の増加の閾値を特定することは不可能かもしれない。このことを明らかにするためには、疫学データが、集団の広範な横断面における化学物質ばく露と疾病リスクの関係を特徴づけ（あるいは前兆病変や重要事象を調べ）、低用量の傾きをしっかりと検討する必要がある。

脆弱な小集団に関連するエビデンスの程度に基づいて、用量反応評価が集団全体に焦点を当てるのか、一般集団と影響を受けやすいサブグループに分けて評価するのかを検討すべきである。集団全体を対象とする場合、従来のアプローチでは、不確実性因子を用いて変動性に対処することになるが、毒性物質に反応して疾患のリスク分布がどのように変化するかを評価することで、変動性のリスクへの影響を解析することも可能である。要するに、不顕性バイオマーカーに基づくリスク分布は、用量反応評価で捉えることができる毒力学的変動の表現である。

別のアプローチとしては、脆弱な小集団を一般集団とは別個のものとして扱い、その集団の実際の用量反応データ、特定の薬物動態または毒物動態学的要因の調整、またはより一般的な補正や不確実性の要因に基づいた用量反応モデルを用いて、その集団に固有の効果を割り当てることである。農薬については、特定の年齢層、疾患（または疾患関連のエンドポイント）、遺伝的変異、または共ばく露が独特の脆弱性を生んでいることが分かっている場合には、一般集団に対する効果の差を推定するよう努めるべきであり、それに基づいて、別個の効果を開発するか、または最も感受性の高い集団または脆弱性のある集団に対する補正を加えた全体の集団に対する単一の効果をベースにすることを検討すべきである。

4.4. ばく露評価の改善

疫学研究における農薬ばく露評価の困難さは、上記したように強調されている。農薬ばく露の記述（特に個々の農薬へのばく露に関する定量的な情報）は、一般的に規制目的のためには十分な詳細が報告されておらず、特に潜伏期間の長い疾患（多くのがんや神経変性疾患など）では、この限界を克服するのは困難である。

ばく露モニタリングを実施するために必要な方法は、申請者が申請書類の中で提出しなければならないことは注目に値する。この規則の要求事項は、ばく露量の判定に使用できる有効な方法を要求している。欧州議会及び理事会の PPP の上市に関する規則（EC）No 1107/2009 に基づき、有効成分のデータ要求を定めた欧州委員会規則（EU）No 283/2013 では、承認前の試験と承認後のモニタリングの両方をサポートするために必要な分析方法に関する情報が記載されている。この文脈では、承認後の要求が最も有用性が高く、規則には実際に次の通りに記載されている。

‘4.2. 承認後の管理及びモニタリング目的のための方法－方法は、以下の目的で提出されなければならない。

a）加盟国が確立された最大残留レベル（MRL）への準拠を決定できるようにするために、6.7.1 項の規定に従って提出されたモニタリング残留物の定義に含まれるすべての成分の決定；これは、植物及び動物由来の食品及び飼料に含まれる、または上述の残留物を対象とする。

b）7.4.2 項の規定に従って提出された土壌及び水の残留基準をモニタリングする目的のための含有全成分の測定。

c）申請者が、作業者、労働者、住民または居合わせただけの者（bystanders）のばく露が無視できる程度であることを示さない場合には、散布中または散布後に生成された有効成分及び関連する分解生成物の大気中の分析。

d）体液中及び組織中の有効成分及び関連代謝物の分析。

これらの方法は、可能な限り、最も簡単な方法を採用し、最小限のコストで、一般的に利用可能な機器を必要とするものとする。分析の特異性を確認し、報告されなければならない。それにより、モニタリング残留物の定義に含まれるすべての成分を測定できるようにしなければならない。必要に応じて、検証された測定法を提出しなければならない。方法の直線性、回収率、精度（再現性）が測定され、報告されなければならない。

データは、LOQ と残留可能性の高いレベル、または LOQ の 10 倍のいずれかで生成されなければならない。LOQ は、モニタリング残留物の定義に含まれる各成分について決定され、報告されなければならない。植物や動物由来の食品や飼料中の残留物や飲料水中の残留物については、方法の再現性は、独立した試験施設のバリデーション（ILV）によって確認され、報告されなければならない。

このことから、要求事項は存在するが、食品や飼料中の残留物のモニタリングよりもヒトのバイオモニタリングの方がやや厳しくないと結論づけることができる。

これらの既存の方法を使用しないと、特定の農薬の規制における疫学的証拠の使用の可能性が制限される。したがって、今後の研究を検討する際には、ばく露の誤分類を避けるために使用する手法を慎重に検討することが重要であり、当該農薬について適切で詳細なばく露評価を行い、適切な用量反応分析を可能にし、使用された方法の妥当性を証明することが重要である。

ばく露は、ばく露される生涯の期間に応じて、異なる健康影響を及ぼす可能性がある。ばく露評価がそのような重要な時期に適切に対応していることを確実にすることで、疾患発症の可能性がある時期のばく露には、より大きな注意を払う必要がある。これは、神経発達、肥満、アレルギー反応など、出生前または出生後早期に発生する複雑な多段階の発達過程を伴う研究に特に関連していると考えられる。このため、単一の期間におけるばく露の測定では、環境因子のすべての健康影響に対する関連ばく露を適切に特徴付けることができない可能性があり、したがって、環境因子に対するいくつかの重要な生物学的に脆弱な期間におけるばく露を測定する必要がある可能性が生じてくる。化学物質の範囲とばく露の強度及び研究範囲ではない他の物質とのばく露も含めた中で、現在のばく露とは異なる可能性がある過去のばく露の評価を構築することは、特に困難である。

農薬ばく露を測定するすべての方法には長所と短所があり、特定の研究計画と目的は、特定の最適なアプローチを通知するために慎重に考慮されるべきである。

観察による研究では、個人レベルのばく露評価は以下の方法を用いて改善することができる。

a）**個人のばく露モニタリング**：これは、測定装置を接触時での農薬濃度を記録するために使用することができる。個人用ばく露モニターは、研究参加者にとって高価で負担が大きいものであった。しかし、技術の進歩により、最近では大気中の浮遊物質ばく露のための個人ばく露モニターが安価で使いやすい装置になり、これらは集団研究に適している。農薬ばく露に特化した個人ばく露モニターには、空気中の濃度を測定するためのセンサー、経皮濃度を測定するための「皮膚」パッチ、他のばく露手段を測定するための塵埃を捕らえる屋内家庭用モニターなどがある。これらのモバイル技術の進歩は、詳細で強固なばく露評価がなされた観察による研究を提供するために利用することができる。このような機器は現在、大規模な集団研究や大規模なコホート研究からのデータを収集するために、ますます適応が増えている。これらは、ライフスタイルやその他の習慣を捉えるための携帯電話や携帯電話アプリケーションを介したリアルタイムのデータ転送などの他の技術的進歩と相まって、次世代の観察による研究では、現在のエビデンスと比較して、はるかに詳細で強固なばく露評価が可能になるだろう。しかし、膨大な量のデータを生成することは、組織的、統計的、技術的な課題、特に追跡調査時間の延長をもたらす可能性がある。倫理や個人データ保護の問題を考慮しなければならず、地域の規制により、このような技術の大規模な使用が妨げられる可能性がある。しかし、このような個人モニターの使用は、異なる潜在的なばく露経路のうちの 1 つの情報を提供するにすぎない。

b）**ばく露のバイオマーカー**（ヒト・バイオモニタリング（HBM））。代替的及び／または補完的なアプローチとして、異なる経路（経皮、吸入及び経口のばく露）を介したばく露の結果である内部ばく露量の確認がある。これらのバイオマーカーは、農薬への総体的なばく露を評価し、累積的なリスク評価に情報を提供する上で重要な役割を果たす可能性がある。バイオモニタリングには、検討対象の化学物質（親化合物または代謝物）またはその病態生理学的影響のマーカー（付加生成物など）の生体試料中の濃度を測定することが必要である。しかし、課題としては、生体試料中の濃度測

定値を関連する用量に外挿することに伴う不確実性が含まれる場合がある。

バイオモニタリングは異物の吸収量を確実に推定できる可能性があるが、現代の農薬とその代謝物は比較的迅速に体内から排泄され、排泄半減期は通常数日で測定される(Oulhote 及び Bouchard、2013 年)。そのため、バイオマーカーの使用はそのため、バイオマーカーの使用は、資源を必要とし、かつ煩わしいものとなります。長期にわたるばく露を監視するために、多数の人を対象に繰り返し実施しなければならない場合には、このプロセスはさらに煩わしいものとなる。

とはいえ、吸収量の正確な積算値を提供できる可能性があるため、農薬とその代謝物の生物学的モニタリングは、ばく露評価の他のアプローチを調整するために有効に利用できる。このようなアプローチの例は、農業健康調査(Agricultural Health Study)で使用されているものである(Thomas ら、2010 年;Coble ら、2011 年;Hines ら、2011 年)。また、長いばく露履歴を構築するために、他の形態のばく露評価と組み合わせて HBM の手法を使用することもできる。

バイオモニタリングは、ばく露の特性評価の精度を向上させ、環境に関連するばく露濃度で発生するばく露の変化を調査することを可能にする。大規模なバイオモニタリング研究で収集されたデータは、今後の疫学研究におけるばく露の分類を支援するための基準範囲を設定するのに有用である。また、バイオモニタリングデータは、リスク評価の改訂を実施するための重要な情報を提供し、有害な健康影響に対して特別なリスクのある小集団を特定するのに役立つ。

生体試料他(Biobanks)は、生物試料の保管場所として、早期ばく露と遅発影響の関係を調査する目的で、ばく露のバイオマーカーを評価するために利用することができる。すなわち、生涯初期に発生したばく露が、生涯後期における疾患(神経行動障害、小児腫瘍、免疫毒性障害など)の発症に重要であるかどうかを調べ、現在の健康ガイドラインに沿って健康リスクを後ろ向きに評価することができる。

尿、血液、毛髪などのヒトの試料中の代謝物レベルの測定結果だけは、実際に受けたばく露量を完全に把握できない。全身または組織・器官の総ばく露量を推定するためには、場合によっては生理学的毒物動態学(PBTK)アプローチを用いた追加評価が必要となる。PBTK モデルは、化学物質の毒物動態を特徴づけるために使用される生理学的なコンパートメントモデルであり、特にヒトにおける化学物質の運命を予測するために使用される。各コンパートメント内で化学物質が受ける血流速度、代謝及びその他のプロセスに関するデータは、PBTK モデルのマスーバランス・フレームワークを構築するために使用される。PBTK モデルは、外部ばく露を体内の内部(標的)ばく露量に変換するだけでなく、バイオモニタリングデータから外部ばく露を推測するためにも使用することができる。さらに、PBTK モデルは検証される必要がある。

毒物動態プロセス(ADME)は、標的に到達した有効成分の「内部濃度」を決定し、この濃度／用量を観察された毒性効果と関連付けるのに役立つ。現行の規制でも試験実施が規定されているが、in vitro 試験、動物試験、ヒト試験など、有効成分の毒物動態に関するすべてのエビデンスを調査することは有益であろう。品質保証の問題や HBM 試験に関連して考慮すべき要素についての更なる議論は、EFSA が委託したプロジェクトの報告書(Bevan ら、2017 年)に記載されている。

ばく露評価は、観察による研究における集団レベルでも、以下のような方法で改善することができる。

a)健康影響や農薬使用に関する**登録データ**や大規模データベース(行政データベースを含む)から得られたデータなど、新しい技術やビッグデータを利用した大規模な疫学研究は、より確かな知見が得られ、最終的には情報に基づいた政策決定や規制に利用できる可能性がある。このようなデータは、維持管理が義務付けられている農家やその他の専門的ユーザーなど、異なる集団による農薬使用の記録が含まれている可能性がある登録データの利用を中心に、多くの努力が必要である[9]。このようなデータは、電子的健康記録(上記参照)にさらにリンクされ、これまでにないサンプルサイズ及びばく露とその後の病気に関する情報を持つ研究を提供し、最終的にはこれまで回答のなかった強固な問

[9] 規則 1107/2009 第 67 条は次のように述べている。記録の保持 1. 植物保護製剤(農薬)の生産者、供給者、流通業者、輸入業者及び輸出業者は、自らが生産、輸入、輸出、保管又は市場に出した植物保護製剤(農薬)の記録を少なくとも 5 年間保管しなければならない。植物保護製剤(農薬)の専門的使用者は、使用した植物防疫製品(農薬)の記録を少なくとも 3 年間、植物保護製剤(農薬)の名称、適用時間及び適用量、植物保護製剤(農薬)が使用された地域及び作物を記載して保管しなければならない。これらの記録に含まれる関連情報は、要求があれば所轄官庁に提供しなければならない。飲料水製造業界、小売業者又は住民などの第三者は、所轄官庁に連絡することにより、この情報へのアクセスを要求することができる。所轄官庁は、適用される国内法または共同体法に従って、当該情報へのアクセスを提供するものとする。

題に答えることができるようになるだろう。同時に、これらの登録や大規模なデータベースでは、有効成分に関する情報をよりよく把握する必要がある。食品中の残留農薬ばく露は、監督下の残留試験と組み合わせて散布日誌データを使用することで、より正確に推定することができる。この方法は、より包括的で強固な源データを含み、使用された農薬をより完全にカバーし、標準的な定量限界（LOQ）以下の残留物のより信頼性が高く正確な推定値を得ることができるという利点がある（Larsson ら、2017 年）。

b）**地理的情報システム**（GIS）や小地域調査への新しい高度なアプローチも、住居ばく露の推定値を提供する追加的な方法として役立つかもしれない。GIS に基づくばく露指標（すなわち、農業用地への住居の近接性や住居周辺への影響のある農地面積）は、検証された場合、バイオモニタリングを補完する有用なツールとなる可能性があり、生物学的半減期の短い農薬へのばく露を評価するために使用されてきた（Cornelis ら、2009 年）。このようなばく露の中には、他の要因の中でも特に風向によって影響を受けるものがあるため、このアプローチを最大限に活用するためには、結果の特別な分析を通じて、この点を考慮に入れる必要がある。また、これらの指標は、特定の指標ではないにせよ、生物モニタリングデータよりも、長期にわたる非残留性農薬への累積ばく露のより代表的な指標となり得る（González-Alzaga ら、2015 年）。

すでに議論したように、個々の化合物の規制リスク評価に有用であるためには、疫学的ばく露評価は特定の農薬に関する情報を提供すべきである。しかし、より一般的なばく露評価を含む疫学的研究は、一般的なリスク因子を特定し、関連するヒト集団における因果関係の推論を示唆する可能性もある。このような観察結果は、全体的な規制政策の情報提供と、さらなる疫学的研究のための事項の特定の両方に重要であるかもしれない。

現代技術の最近の進歩により、新しい分析方法を用いて農薬ばく露を前例のない範囲で推定することが可能になった。

a）メタボロミクス（metabolomics）やアダクトミクス（adductomics）などのいわゆる**オミクス技術**の発展は、生物学的マトリックス（血液、唾液、尿、毛髪、爪など）中に経時的に記録された異物や代謝物から、DNA やタンパク質との共有結合体（アダクトミクス）や生物学的経路の理解に至るまで、幅広い分子の測定を通じてばく露評価を改善するための魅力のある可能性を提示している。これらの方法論は、他のツールと組み合わせて使用することができる。また、このような技術を規制毒性学に応用するには、さらなる研究が必要であるという認識と関心がある。エクスポソーム（exposome）（一生の間に個人が受けたばく露の全体）の利用は、「オミックス」技術とヒトのバイオモニタリングに適したバイオマーカーを使用することで、より良い結果を得ることができるかもしれない。それにもかかわらず、これらの方法論の検証が不足していることと、大規模での使用を制限するコストのため、重要な制限が認められなければならない。

b）環境ばく露は従来「一回のばく露に一回の健康影響」というアプローチで評価されてきた。これに対して、**エクスポソーム**は、受胎以降のヒト環境ばく露の全体を網羅しており、遺伝学の知識を補完することで、疾患の病因における環境要因をよりよく特徴づけることができる。このように、エクスポソームには、生涯にわたる化学的ばく露だけでなく、感染症、身体活動、食事、ストレス、内部生物学的因子（代謝因子、腸内フローラ、炎症、酸化ストレス）などの外部環境因子や内部環境因子も含まれている。完全なエクスポソームを構築するためには、生涯にわたって継続的に異なる源からの多くの外部ばく露と内部ばく露を統合しなければならない。しかし、真に完全なエクスポソームを測定することは不可能である。エクスポソームのこれらすべての領域を従来のものとは異なるアプローチで捉える必要があるが、この目的のためには単一のツールでは十分ではないと考えられている。

ばく露のより総合的なアプローチは、現在の疫学研究における従来の「一回のばく露－一回の健康影響」アプローチに取って代わることを意図したものではない。しかし、それは複雑で多因子性の慢性疾患の予測因子、リスク因子、保護因子についての理解を向上させるものである。エクスポソームは、生涯にわたる環境の影響やばく露を総合的に記述し、統合する枠組みを提供している（Nieuwenhuijsen、2015 年）。

これらの潜在的なバイオマーカーを検証し、最終的にはばく露評価の改善につなげるためには、共同研究や、大規模なコンソーシアムを形成する疫学研究や探索的研究の統合が必要である。エクスポソームパラダイムを従来のバイオモニタリング手法に組み込むことは、ばく露評価を改善する手段となる。エクスポソーム拡大関連研究（EWAS）は、健康な人と病気の人の血液中の何千もの化学物質を測定し、病気との関連性を検査し、ばく露源を特定し、作用機序を確立し、因果関係を明らかにするために、その後の調査で対象とすることができるばく露の有用なバイオマーカーを特

定することを可能にする(Rappaport、2012 年)。これらの主要な化学物質を特定し、症例と対照の独立したサンプルで疾患との関連性を検証した後、これらの化学物質は、大規模な集団からの血液を対象とした分析において、ばく露または疾患進行のバイオマーカーとして使用することができる。

エクスポソームの概念に関連して、オミクス技術は複雑な化学物質の混合物への累積ばく露に対する生物学的反応の特性やシグネチャーを測定する可能性を持っている。重要な進歩は、特定の生物学的試料中の個々のばく露を個別に評価することなく、エクスポソームを特徴づけることができるユニークな生物学的マトリックスを特定することであろう。オミクスデータの非標的特性は、ばく露に対する生物学的反応をより全体的な方法で捉え、ばく露に関連した健康影響を裏付けるメカニズム論的な情報を提供することになる。重要なことは、オミクスツールは、多様なばく露がどのようにして共通の経路で作用し、同じ健康影響を引き起こす仕組みを明らかにすることができるということである。

改良されたばく露評価は関連性を検出する力を高めるが、どのような個々の研究においても、各被験者のばく露評価を実施するために使用する供給源と被験者の総数のバランスを最適化することにより、研究の全体的な力を最大化することが必要である。

4.5. 健康影響

農薬については、これらの化学物質が単一の疾患領域に関連して特定の効果を示していないため、健康影響は広範囲にわたる。それぞれの健康影響について、文献には複数の定義が存在していて、程度の差こそあれ異なるデータベース間での再現性は不明であり、一般化できないという制限がある。健康影響の適切な定義は、観察による疫学研究の妥当性と再現性にとって非常に重要であり、これらの定義の一貫性と明確さは研究間で考慮する必要がある。前向きな観察研究では、明確な健康影響の定義、包含基準と除外基準、標準化されたデータ収集があるが、後ろ向き研究では通常、主にコード化されたデータに基づいた健康影響の同定に頼っており、疾患の分類とコード化は時間の経過とともに変化する可能性がある。主要な健康影響を定義するために使用された実際のコードの詳細な記述と検証作業の結果は、今後の研究活動にとって貴重なものである(Stang ら、2012 年;Reich ら、2013 年)。コード化された疾患の例としては、例えば ICD-10 があり、これは広範囲の悪性疾患を標準化するためのツールとして使用することができる。

いくつかのサーベイランス研究では、すべての潜在的な症例を特定するために感度の高いより広い定義を使用し、その後、偽陽性の数を減らし、結果としてより正確な症例を得るために、高い陽性予測値のより狭く、より正確な定義を適用することが望ましいとされている。対照的に、正式な疫学研究では、特定のイベントの定義が使用され、その精度を決定するために検証される。しかしながら、「検証」では新たな定義をテストしないので、感度や特異度を測定できないでだろう。

代替エンドポイントは、有効性が確認されていない限り避けるべきである。代替健康影響の妥当性を評価する基準には、以下のようなものがある。

- 代替指標が疾患の原因経路内にあることが示されていること。これは以下のエビデンスによって裏付けられる:バイオマーカーの反応が病理学と相関しており、他のバイオマーカーと比較して性能が向上していること;生物学的な理解と毒性との関連性(反応のメカニズム);メカニズム的に異なる化合物に対する一貫した反応と性、系統、種の違いによる類似した反応;用量反応の存在と反応の大きさと時間的関係;毒性に対する反応の特異性;すなわち、バイオマーカーは他の組織の毒性に対する反応や、標的臓器の毒性を伴わない生理学的効果を伴わない生理的効果を反映してはならない。
- 代替健康影響と真の健康影響の両方を使用した少なくとも 1 つのよく実施された試験があること(Grimes 及び Schulz、2005 年;Ia Cour ら、2010 年)。これらの基準を評価するために、いくつかの統計的手法が使用されて、それらが満たされていれば、代替指標の妥当性が高まる。しかし、多くの場合、不確実性が残っているため、疫学研究に代替指標を適用することは困難である(Ia Cour ら、2010 年)。

EU 全体の健康影響に関するデータは非常に広範囲に及ぶ可能性がある。これを効果的に管理することができれば、非常に大規模なサンプルサイズを用いて悪影響を評価する疫学研究において、より大きな統計力を発揮できる可能性がある。これらの研究に必要な前提条件は、新たな軽微な影響、慢性的な影響、または層別化した場合の小集団

への影響を検出する可能性があるが、リスク評価の範囲を超えている。これらの研究には、調和のとれた診断、データ保存及び社会的利益のための匿名化された個人データへの法的に承認されたアクセスと相まって、健康情報学への国境を越えたアプローチが含まれている。健康記録には、適切な中毒症候群の分類が含まれていなければならない。後者は、入力データの品質を保証するために、医療と医療補助のトレーニングの改善を必要とするかもしれない。

生物学的モニタリングが採用されるもう一つの機会は、調査がいわゆる影響のバイオマーカーを含む場合である。これは、定量化可能な生化学的、生理学的、またはその他の変化であり、その大きさに応じて、確立された、または可能性のある健康障害や病気に関連している。影響のバイオマーカーは、機能的または構造的損傷に先行する初期の生化学的変化を反映している必要がある。このように、最終的に毒性につながるメカニズムの知識は、特定の有用なバイオマーカーを開発するために必要であり、その逆もまた然りで、影響のバイオマーカーは、疾患の発生のメカニズムの経路を説明するのに役立つかもしれない。このようなバイオマーカーは、生物学的システムにおける初期の可逆的な事象を特定するものであり、後の反応を予測するものでなければならず、その性質上、前臨床的なものと考えられる。実験的ーオミクス技術の進歩は有望であり、リスク評価戦略、すなわち作用機序、反応バイオマーカー、内部ばく露量の推定、用量ー反応関係に関する確かな情報を提供するだろう（DeBord ら、2015 年）。これらの技術は、その妥当性と信頼性を評価するために検証されなければならない。妥当性が確認されれば、それらの技術は規制目的で利用できるようになる。

5. 農薬リスク評価への警戒データの貢献

第 2-4 節で議論した正式な疫学調査に加えて、その他のヒトの健康データは、その場限りの報告書から、あるいは計画的なプロセスとして、すなわち公衆衛生当局や認可者によって国家レベルで実施されているモニタリングシステムを通じて、生成することができる。第 2-4 節に沿って、本節ではまず、このようなモニタリングシステムがどのように運用されるべきか、農薬のモニタリングに関する現状はどうなっているのか、そして改善のためにどのような勧告ができるのかをレビューする。

5.1. ケースインシデント研究の一般的な枠組み

有害事象の収集、報告、評価を継続的に行うことは、同じ有害事象が後から別の場所で発生する可能性を減らすことで、利用者やその他の人々の健康と安全の保護を向上させ、また、そのような事象の結果を緩和する可能性がある。そのためには、当然ながら、収集した情報をタイムリーに発信する必要がある。このようなプロセスを警戒（vigilance）と呼んでいる[10]。

例えば、EU では、医薬品の安全性監視は医薬品安全性監視（pharmacovigilance）して知られており、医薬品安全性監視システムは、加盟国の規制当局、欧州委員会、欧州医薬品庁（EMA）の間で運営されている。一部の加盟国では、国内の管轄当局の調整の下に地域センターが設置されている。製造業者や医療従事者は、国レベルの管轄当局に事件を報告する。これにより、有害事象に関するあらゆる情報が記録され、一元的に評価され、その後の対応について他の当局に通知することができる。記録は EMA によって一元化され、欧州の医薬品安全性監視システムの調整をサポートし、医薬品の安全で効果的な使用に関するアドバイスを提供する。

5.2. ケースインシデント報告の現在の枠組みの主な限界

いくつかの EU の規制では、ヒトに農薬が原因で発生した有害事象（職業環境での急性または慢性ばく露後に発生したもの、偶発的または故意の中毒など）の通知及び／または収集及び／または報告を義務付けている。これらには以下のものが含まれる。

・ EC 規則 1107/2009 の第 56 条は、「植物保護製剤（農薬）の認可を受けた者は、直ちに加盟国に通知しなけれ

[10] 調査という概念は、何かを測定し記録するための単一の努力を意味し、サーベイランスとは、疾病の不在を証明したり、疾病の存在や分布を特定して情報を適時に発信できるようにするために、集団の傾向を検出するために、標準化された調査を繰り返すことを意味する。モニタリングとは、集団の環境や健康状態の変化を検出するために、日常的な測定や観察を断続的に分析することを意味するが、反応を引き出すことはない。監視は、綿密かつ継続的に注意を払うプロセスを意味するため、監視や単なるモニタリングとは異なり、この背景では特に化学物質の使用に関連した販売後の事象を扱う。

ばならない」と規定している。この目的のために、認可保有者は、植物保護製剤（農薬）の使用に関連して、ヒト、動物及び環境におけるすべての疑われる有害な反応を記録し、報告しなければならない。通知義務には、国際機関や第三国の植物保護製剤（農薬）や有効成分を認可する公的機関による決定や評価に関する関連情報も含まれていなければならない。

- 農薬の持続可能な使用を達成するための共同体行動の枠組みを確立した EC 指令 128/2009 の第 7 条は、次のように要求している。加盟国は、作業者、農業労働者、農薬散布地域の近くに住む人など、定期的に農薬にばく露される可能性のある集団の間で、農薬による急性中毒事故や慢性中毒の発生状況に関する情報を収集するシステムを設置しなければならない。3.3.情報の類似性を高めるために、欧州委員会は加盟国と協力して、2012 年 12 月 14 日までに「農薬使用がヒトの健康と環境に与える影響のモニタリングと調査に関する戦略的ガイダンス文書」を作成する。しかし、この意見書を発表した時点では、この文書はまだ公表されていない。

間接的ではあるが、農薬と報告に適用される追加の規制が 3 つある。

- 農薬の統計に関する EC 規則 1185/2009 は、加盟国が調和のとれたフォーマットに従って農薬の販売と使用に関するデータを収集することを要求している。上市に関する統計は毎年欧州委員会に、農業利用に関する統計は 5 年ごとに送信されなければならない。
- 食品法の一般原則と要件を定めた規則（EC）178/2002 の第 50 条では、食品と飼料を対象とした改良・拡大された迅速警報システム（RASFF）が設定されている。このシステムは欧州委員会によって管理されており、ネットワーク加盟国、欧州委員会、当局が加盟している。それは残留農薬の認可されていない事例や食中毒の事例を報告する。
- EC 規則 1272/2008（CLP 規則）第 45 条（4）：EU 加盟国の市場に危険な化学物質の混合物を市場に出す輸入業者と顧客ユーザーは、その加盟国の任命機関／毒物センターに通知書を提出しなければならない。通知書には、化学成分や毒物学的情報、混合物が属する製品カテゴリーなど、混合物に関する特定の情報を記載する必要がある。通知書に製品分類に関する情報を含めることで、指定団体／毒物取締センターは、同等の統計解析（例えば、リスク管理措置の実施）を行い、報告義務を果たし、MS 間で情報交換を行うことができる。したがって、製品カテゴリーは実際の緊急時の医療対応には使用されないが、ばく露や中毒の傾向を特定し、将来の中毒事例を防ぐための対策をとることができる。正式に採択された場合、新規則は 2020 年 1 月 1 日から適用される。

実質的な立法規定がある一方で、今日までのところ、医薬品安全性監視システムに類似した単一の EU「植物薬理監視」[11]システムは PPP には存在しない。むしろ、加盟国の公衆衛生にリスクをもたらす可能性のある化学物質のハザードについて警告、通知、報告、情報共有を行うための多くの警告システムが EU 内で開発されている。これらのシステムは、医薬品、食品、消費者製品、労働災害、国際保健規則（IHR）に基づく通知、EU 毒物センターや公衆衛生当局によって検知した事象など、さまざまな分野をカバーしている。これらのシステムのそれぞれは、管轄官庁、公的機関、政府、規制当局、公衆衛生当局にタイムリーな警告を通知し、配布して、公衆衛生へのリスクを最小限に抑え、管理するための効果的な行動をとることを可能にしている（Orford ら、2014 年）。

EU では、急性農薬ばく露・事故に関する情報は、主に毒物管理センター（PCC）によって収集・報告されたデータに基づいている。PCC は、一般集団や職業環境において、自分たちが知っている急性と慢性のばく露／中毒の両方の事例を収集している。通常、症例は十分に文書化されており、情報にはばく露・事故の状況、原因物質と疑われるものの説明、ばく露のレベルと期間、臨床経過と治療、因果関係の評価が含まれている。重症の場合は、通常、血液や尿中の毒素や代謝物の測定が行われる。しかし、センターに報告された症例の追跡調査は、長期化する可能性のある影響を特定するために、さらに注意を払う必要がある。

毒物センターのデータを使用するには、2 つの重要な障害がある：各国の毒物センターからの報告書は常に公表さ

[11] 「フィトビジランス（phytovigilance）」は植物に対する警戒システムを意味し、農薬は作物の「薬」であることを意図しているため、ここでは「フィトファーマコビジランス（phytopharmacovigilance）」という用語がより適切であると考えられている。さらに、フランスでは、土壌、水、大気、環境、動物のデータなどをカバーする広い用語として使われている。

れているわけではなく、公表されている場合でも、データ収集の形式やコーディング、因果関係の評価には大きな不均一性がある。実際、各加盟国は独自の収集活動のためのツールを開発しており、ばく露データの比較や交換には困難が伴う。2012 年、欧州委員会は、新たな化学物質事象への欧州の対応を支援するための共同研究開発プロジェクト「化学物質健康脅威のための警告・報告システムフェーズ III(ASHTIII)プロジェクト」に資金を提供した。検討された様々なツールや方法論の中で、欧州の PCC からのばく露データを交換・比較する方法が開発された。実現可能な研究として、ワークパッケージ 5 には、加盟国が農薬ばく露データを比較できるようにするための、調和のとれた強固なコード化システムの開発が含まれていた。しかし、PCC コミュニティとの協議の結果、データのコーディングと収集活動のさらなる調整が必要であることが示された。その結果、ばく露データを加盟国間で比較できるようにするためには、EU と加盟国レベルでの更なる支援と調整が必要であると結論づけられた(Orford ら、2015 年)。

PCC が収集したデータに加えて、いくつかの加盟国は、労働衛生監視に特化したプログラムを立ち上げた[12]。これらは、業務上の農薬による傷害や病気、中毒が疑われる症例について、医師による自発的なイベント通知(使用者による自己申告の場合もある)に基づいている。医療データに加えて、収集された情報には、作物の種類、散布方法、温度、風速、個人用保護具の着用状況などに関するデータが含まれる。一度収集されたこれらのデータは調査され、定期的に報告書が発行され、再登録中の製品の安全性を評価するための有用な情報となる。これらのデータはまた、新たな問題を浮き彫りにし、政策立案者のためのエビデンスに基づく予防措置を策定することを可能にする。EU レベルでは、欧州労働安全衛生庁(EU-OSHA)[13]は、職業上の農薬関連疾病データのモニタリング方法をほとんど持っていない。米国では、国立労働安全衛生研究所(NIOSH)が資金を提供し、農薬に特化したプログラムがいくつかの州で実施されている[14]。

要約すると、現在、ヒトのデータは、症例報告書や症例集積、毒物センターの情報、検視官の裁判結果、労働衛生監視プログラムや市販後の監視プログラムの形で収集されている。しかし、申請者が提出した医療データには、このような情報がすべて含まれているわけではない。これは、さまざまな情報源が多様で異質な性質を持っているため、アクセスできないものもあるためである。

- 工場生産労働者の労働衛生監視を通じて収集されたデータ、あるいはそれが行われたとしても、医療データは非常に限られており、一般的には基本的な臨床血液測定、身体検査、潜在的にはどこでどのようにばく露されたかという単純な指標であり、通常は長期的なフォローアップは行われていない。さらに、最新の工場(特に EU)での労働者のばく露は一般的に非常に低く、多くの場合、潜在的なばく露は(特定の化学物質に特化した施設でない限り)様々な農薬へのばく露である。
- さらに、製造中の有効成分への職業上ばく露からのデータの報告は、しばしば調合された植物保護製剤(農薬)との接触から生じる観察結果と組み合わされる。実際、植物保護製剤(農薬)中の共配合剤の存在は、急性毒性学的プロフィルを変更することができる。したがって、適切な評価を容易にするために、ヒトで収集した結果を報告する際には、それ自体が有効成分なのか PPP なのかを明確に特定しなければならない。

EC 規則 283/2013 の第 5.9 章の一部でもある有効成分や配合された植物保護製剤(農薬)や提案された治療法による中毒の診断に関する特定のデータの要求に関しては、情報が欠落していたり、毒性作用のモードがヒトで起こることが知られていて特定の解毒剤が特定されている場合に限定されていたりすることがよくある。

12 例えばフランスの Phyt'attitude は、Sociale Agricole, Mutualite, Sociale Agricole によって開発された警戒プログラムである: http://www.msa。

13 https://osha.europa.eu/en/about-eu-osha

14 SENSOR プログラム: https://www.cdc.gov/niosh/topics/pesticides/overview.html

5.3. ケースインシデント報告の現行枠組みの改善提案

重複と努力の無駄を避けるために、論理的な次のステップは、すべての関係する公的部門と民間部門の関係者と一緒に、医薬品のために実施されているものと同様の化学物質のための EU の「植物薬理監視」システムを開発することであろう。このネットワークは、献身的で特別な訓練を受けた地方の産業保健医や開業医を基盤とすることができ、システムを確立し、成功裏に維持するために加盟国が供給源を配分すべきである。実際、このようなネットワークは急性の影響を検出するのに有用であろう。また、特定の健康影響(喘息、感作など)や新たな職業関連疾患の検出のためのセンチネルサーベイランスネットワークとしても機能するであろう。実際、このようなシステムを段階的に構築する方法については、すでに多くの経験が得られているが、それにもかかわらず、これが実施されるまでには何年もかかることが想定されている。収集されるデータの特性(情報源は多様である可能性がある)、収集された情報の質と完全性(特に状況)、観察された効果の重症度と説明責任(観察された効果と製品との間のリンク)など、いくつかの困難が生じる。ルールは、ある「評価者」から別の「評価者」まで同一であるように定義されなければならない。植物薬理監視システムが目的に完全に適合していることを保証するために、ネットワークは長期的に安定していなければならない(例えば、関与する国の組織の継続性、採用された一貫した方法論など)。植物薬剤モニタリングデータの使用は、リスク評価の目的に限定されることはなく、リスク管理上の政策決定(例えば、製品認可の条件の改定や最終的には製品の取り下げなど)に影響を及ぼす可能性があるが、これは最初からすべての利害関係者に明確でなければならない。

このようなシステムは、(主に)農薬として使用されている化学物質のみを対象としたものでは意味がないかもしれない。しかし、すでに農薬に関する法律の規定があることを考えると、このシステムの開発は農薬に優先して行われる必要があるかもしれない。

結論として、欧州委員会は加盟国とともに、EU 全体の農薬の警戒枠組みの開発に着手すべきである。これには以下が含まれるべきである。

- EU レベルでのヒトでの健康影響データ収集活動の調和
- EU 全体のデータベースの編集の調整
- 各加盟国で発生したすべての PPP 中毒を収集するために、各国レベルでのポイズンセンターと規制当局との連携を改善すること
- 因果関係のデータ評価の整合化を伴う農薬使用がヒトの健康に及ぼす影響のモニタリングに関するガイダンス文書
- EU 全体を対象とした定期的な報告書

6. 農薬のリスク評価を支援するための疫学研究と監視データの利用の提案

本節では、実験的研究に基づくリスク評価プロセス(第 6.1 節)を概説し、そのプロセスに疫学的研究がどのような情報を付加しうるかを論じる。次に、第 6.2 節では、疫学研究の信頼性の評価について述べる。6.3 節では、信頼性があると認められた 1 つ以上の研究の関連性を評価する。

6.1. リスク評価プロセス

リスクアセスメントとは、健康に悪影響を及ぼす可能性のある化学物質やその他の汚染物質、薬剤によるヒトや環境へのリスクを評価するプロセスである。規制目的のために、リスク管理者に情報を提供するために使用されるプロセスは、4 つのステップで構成されている(EFSA、2012 年 a)。一方では、毒性影響の特性(ハザード同定)と、農薬と毒性影響の間に考えられる用量反応関係(ハザード特性評価)に関する情報が収集される。一方で、ヒト(消費者、散布者、労働者、居合わせただけの者、住民)と環境へばく露可能性についての情報が求められる(ばく露評価)。これら 2 つの要素は、集団が基準ばく露量を超える量にばく露される可能性があることを推定するために、リスク特性評価の中で考慮される。通常、これは規制目的のためのリスク管理者への情報提供に用いられる。

a)ステップ 1. *ハザードの同定*

疫学的研究と監視データは、農薬ばく露と健康影響とが関連する可能性を示すことができるため、ハザードの特定に関連している。この背景では、疫学的データは、実験モデルでは検出されなかった影響を「ホライズン・スキャニング」

する上で、非常に貴重な情報を提供することができる。重要なことは、これらの研究はまた、脆弱な集団のサブグループ、生涯における感受性の高い時期、性別による選択的影響など、リスクが高まる可能性についての情報を提供することである。

b)ステップ 2. *ハザードの特性評価*(用量反応評価)

前述したように、通常、疫学的データを使用する場合には、ばく露量が割り当てられることはほとんどないため、古典的な用量反応の枠組みは考慮されない。質の高い疫学研究が利用可能な場合の課題は、それらを数値入力としてスキームに統合するのが最善かどうかを見極めることである。農薬のリスク評価に疫学的データを使用する場合、用量反応フレームワークが考慮されることはほとんどない。しかし、EFSA CONTAM パネルのこれまでの科学的見解では、基準ばく露量を設定するための基礎として疫学を使用してきた、特にカドミウム、鉛、ヒ素、水銀の場合は、最もよく知られていてデータが豊富である(EFSA、2009 年 a,b、2010 年 b、2012 年 b)。これらが用量反応評価の基礎とならない場合でも、監視と疫学的データは、実験動物を用いた用量反応研究の妥当性を検証したり、無効にしたりするための裏付けとなるエビデンスを提供することがある。化学物質の様々な用量とばく露された集団における有害な影響の発生率との間の関係を特性評価するためには、ばく露または用量の特性評価、反応の評価、無影響量を特定するために観察されたデータが適合する用量反応モデルの選択が必要である。2 つの課題が提起される。すなわち、無影響量を特定するために、疫学的データから用量反応を導き出すことができるのか、ということである。もしそうでない場合、疫学的情報はハザードの特性評価に貢献できるのか、ということである。

用量反応関係を理解することは、EU の優れた植物保護対策が予想されるよりも高いばく露量の使用による有害な健康影響が関連していることが証明されるが、低いばく露量の使用では関連性が観察されない。この背景では、RR または OR を明らかにした疫学研究の統計的要約は、研究デザインが必要な基準を満たしている場合には、ハザード特性評価プロセスに投入するための有用な定量的情報となる可能性があることは明らかである。

c)ステップ 3. *ばく露評価*

ばく露の評価に関するデータは、制御されていない様々な「実社会」の要因が解析を混乱させる複雑な状況では、推定が困難なことが多い。前述したように、現代の生物学的モニタリングは、コストの高さ、実施可能性、ロジスティックスなどの実際的な理由から、一般のヒト集団ではほとんど実施されていない。しかし、近い将来、農薬への定量的ばく露に関する生物学的モニタリング研究やデータが増加することが予想されている。

ステップ 4. *リスクの特性評価*。この最後のステップでは、ばく露に関するデータを健康ベースの基準値と比較し、ばく露された集団における健康障害のリスクを推定する。ヒトのデータは、標的臓器、用量反応関係、毒性影響の可逆性に関する完全な毒性学的データベースからの外挿に基づいて行われた推定の妥当性を検証するのに役立ち、基準値の定義に直接影響を与えずに外挿のプロセスを再確認するのに役立つ(London ら、2010 年)。

疫学的データは、不確実性因子(UF)との関連で考慮されることもある。一般的に動物データでは、影響の種間変動を考慮するために 10 の UF が使用され、これにさらに 10 の係数を加えてヒト集団の異なる部分の感受性の変動を考慮する。しかし、ヒトのデータのみを考慮する場合(動物のデータよりも重要な場合)もあり、種族間のばらつきを考慮した 10 の係数が適用される。現時点では、規則(EC)No 1107/2009 の第 4 条(6)が次のように規定していることに留意する。「ヒトの健康に関連して、ヒトから収集したデータがない場合、動物試験に由来する安全マージンを低下させてはならない」と規定している。このことの意味するところは、リスク評価において疫学的データはリスク評価で使用される警戒レベルを高めるためにのみ使用され、関連するヒトのデータが入手可能であっても UF を低下させるために使用されてはならないということである。

6.2. 個々の疫学研究の信頼性の評価

WOE 評価のために疫学をどのように考慮すべきかを決定する際に考慮すべき因子は以下に記載されており、観察疫学的研究のためのバイアスのリスクツールで広く概説されている[15]。以下の例は、網羅的なリストではないが注目すべ

[15] 介入またはばく露の観察研究におけるバイアスと交絡因子のリスクの評価。RTI アイテムバンクのさらなる発展(https://www.ncbi.nlm.nih.gov/books/NBK154464/)とコクランハンドブック。

き因子を示している。

- 研究デザインと実施。研究デザインは、ばく露と健康影響及びリスクのある集団の予想される分布を考慮して適切なものであったか？その研究は主に仮説生成モードまたは仮説検証モードで実施されたか？
- 集団。研究は、十分に定義された集団から目的の個人をサンプリングしたか？研究は、ばく露群と非ばく露群の健康影響について有意な差を検出するのに十分な統計力と精度を有していたか。
- ばく露の評価。ばく露の評価に使用された方法は有効で、信頼性があり、適切であったか？広範囲のばく露が調査されたか？ばく露は定量的レベルで評価されたのか、それともカテゴリカルまたは二分法（例：経験対未経験）で評価されたのか？ばく露は前向きに評価されたか、あるいは後ろ向きに評価されたか？
- 健康影響の評価。健康影響の評価に使用された方法は有効で、信頼性が高く、適切であったか？健康影響に関するデータ収集には標準化された手順が用いられていたか。情報の偏りを避けるために、健康影響はばく露状態とは独立して把握されていたか？
- 交絡因子の管理：潜在的な交絡因子が適切に特定され、考慮されていたか？それらはどのように管理されていたか？これらの因子を記録するために使用された方法は有効で、信頼性があり、適切であったか？
- 統計解析。研究は、対象となる健康影響に対するばく露の独立した影響を定量的に推定したか？データの解析において交絡因子が適切に管理されていたか。
- 研究の報告は適切であり、透明性の原則とSTROBE声明（または同様のツール）のガイドラインに従っているか。

研究の評価は、それぞれの研究が持つ可能性のある潜在的な限界の特性と、疫学的データベースの全体的な整合性の評価を示すものでなければならない。

さらに、他の既知のリスク因子に関する健康影響の特性と特異性は、リスク評価目的のためのヒトデータの評価に影響を与える可能性があり、特に誘発期間や潜伏期間の長い慢性的な影響のような複雑な健康エンドポイントの場合には、その評価に影響を与える可能性がある。

表 2 は、単一の疫学研究で評価すべき主なパラメータと、各パラメータの関連する程度（低、中、高）を示している。特定の科学的考察はケースバイケースで適用されるべきであるが、これらの基準を厳格かつ明確な方法で実施することは非現実的である。

表 2：疫学的観察研究の重み付けのための研究の質に関する考察[a]、[b]

パラメータ	高	中	低
試験のデザインと実施	前向き研究 特定の仮説（化合物と健康影響の特定）	症例対照研究。ばく露または健康影響評価を十分にカバーしていない前向き研究	横断、生態学的研究 症例対照研究では、ばく露や健康影響評価が十分にカバーされていない
集団	ランダムサンプリング。十分な検出力を保証するのに十分な大きさのサンプルサイズ 母集団の特性が十分に把握されている（脆弱なサブグループを含む）	疑わしい研究検出力、詳細に正当化されていない 標的集団の代表的なサンプルではない 母集団の特性が十分に解明されていない	研究集団の選定方法についての詳細な情報がない 母集団の特徴が十分に解明されていない
ばく露評価	検証された方法を用いた正確かつ精密な定量的ばく露評価（ヒトのバイオモニタリングまたは外部ばく露） 被験者が回答した化学物質ばく露に関する有効なアンケート及び／またはインタビュー	特定のマトリックス中の非有効なサロゲートまたはバイオマーカーと外部ばく露 被験者または代理人が回答した化学物質ばく露に関するアンケート及び／またはインタビュー	乏しいサロゲート 質の低いアンケート及び／またはインタビュー；化学物質のグループについて収集された情報 化学物質に特化したばく露情報は収集されていない；農薬の使用の有無の一般的な評価
健康影響評価	有効で信頼性の高い健康影響評価 研究集団において標準化され、妥当性が確認されていること カルテまたは診断結果が記載され	標準化された健康影響、母集団で有効性が確認されていない、またはスクリーニングツール、またはカルテが不明確	標準化されていない、検証されていない健康影響 不適切な健康影響、または自己報告された健康影響

	ていること	な場合	
交絡因子コントロール	科学的な課題に関連する重要な交絡因子と標準的な交絡因子の適切なコントロール 明らかに示された交絡因子を慎重に考慮	交絡因子は部分的にコントロール 交絡因子と標準変数を中程度にコントロール 科学的な課題に関連するすべての変数が考慮されているわけではない	研究のデザイン及び解析段階で潜在的な交絡因子及び効果修飾因子をコントロールしていない
統計解析	研究デザインが適切であること、適切なサンプルサイズに支えられていること、データを最大限に利用していること、よく報告されていること（選択的ではない） 交絡因子をコントロールするための統計的手法が用いられており、調整済み及び未調整の推定値が提示されている サブグループと相互作用解析が実施されている	受け入れ可能な方法、情報を失う分析的な選択、明確に報告されていない 事後分析を実施したが、明確に示された	記述的統計、または二変量分析の疑わしいものだけが作られている 比較が行われていない、または明確に記載されていない 分析の不備（多変量解析など）
報告	材料と方法の主要な要素と結果は、十分に詳細に報告されている 研究の各段階における参加者数が報告されている 調査中の関連性のについて信憑性のあるメカニズムが示されている	材料と方法のいくつかの要素や結果は、十分な詳細が報告されていない 結果の解釈は中程度に対応	報告の不備（効果推定値の解釈、交絡因子コントロール） 選択的報告 ばく露と健康の関係に影響を及ぼす可能性のある関連因子に関する情報の不足 推論目的の焦点がずれている 正当化された結論ではない

（a）：パラメータ全体の総合的な評価に基づく総合的な研究品質ランキング。
（b）：Muñoz-Quezada ら（2013）と LaKind ら（2014）を順に引用した US-EPA（2016）からの引用。

上記の評価が、疫学研究が評価され定量的にまとめられているエビデンス総合演習の一部である場合、農薬ばく露に関連する絶対的リスクをより正確に推定し、さらに定量的なリスク評価を行うことが可能となる。

農薬疫学データの場合には、バイアスのリスクと信頼性に関してヒトデータを整理するための第一段階として、3 つの基本的なカテゴリーが提案されている[16]。(a)バイアスのリスクが低く、信頼性が高い（上記の品質要因のすべて、または大部分が軽微な方法論的限界で対処されている）；(b)バイアスのリスクが中程度で信頼性が中程度（上記の品質要因の多くが中程度の方法論的限界で対処されている）；(c)バイアスのリスクが高く、信頼性が低い（結果の妥当性を低下させる、または潜在的な因果関係をほとんど解釈できない、といった重大な方法論的限界や欠陥があるため）。後者の研究は、主にばく露評価の不備、ばく露及び／または健康影響の誤分類、または関連する交絡因子の統計的調整の欠如により、リスク評価には受け入れられないと考えられている。リスク評価は、十分に定義されたデータ品質基準を満たしていない疫学研究の結果に基づくべきではない。さらに、予備的研究の結果は、リスク評価に使用する前に、将来の研究で確認する必要がある。

6.3. 疫学研究のエビデンスの強さの評価

このセクションでは、農薬とヒト健康影響との関連性に関する様々な疫学研究から得られた結果を組み合わせて要約することに関連したいくつかの重要な問題について簡潔に論じている。

疫学研究の重み付けのアプローチは、主に修正された Bradford Hill 基準に基づいている。これは、事象と起こりえる結果（強さ、一貫性、特異性、時間性、生物学的勾配、妥当性、統一性、実験と類推）との間の潜在的な因果関係を示すエビデンスを提供する条件のグループである（表 3）。明らかに、これらの基準を満たせば満たすほど、意味のある

[16] これらのカテゴリーは、現在 EFSA が農薬有効成分のピアレビューに使用している、許容可能、補助的、非許容のカテゴリーに準拠している。

関連性の証拠としてその関連性を提起する根拠が強くなる。しかし、Bradford Hill は、因果関係とは何かを明確にすることを目的とせず、基準を十分に、あるいは絶対的に必要なものとしてかんがえず、単に常識的な評価の中で考慮することが重要であると考えている。

表 3:エビデンス統合のための修正された Bradford Hill 基準に基づく WOE 解析の考察

カテゴリー	考察事項
関連性の強さ	関連性の強さ（関連性の大きさだけでなく、統計的有意性も）の評価には、基礎となる方法の検討、文献の WOE との比較及びここで議論されている他の基準を含む他の背景的要因の考慮が必要である。
関連性の一貫性	関連性は、複数の独立した研究、特に異なる計画で、異なる状況下で異なる集団で実施された研究において一貫性があるべきである。この基準は、現代のデータ統合に照らして、すべてのエビデンス系統（疫学、動物実験、in vitro システムなど）にまたがって一貫性のある結果にも適用される。
特異性	特定の結果をばく露に結びつけるエビデンスの独自の基準は、因果関係についての強力な議論を提供できるようになり、データ統合の背景の中で新たな興味深い意味合いを持つようになったかもしれない。データの統合は、複雑なばく露に関連した様々な結果の中から、いくつかのメカニズム論的な特定性を明らかにするかもしれない。特異性の欠如は、疾患に関連する特異的な薬剤を絞り込むのに役立つかもしれない。
時間性	薬剤へのばく露と適切な時間枠内での影響の出現との間の時間的順序のエビデンスは、因果関係を支持する最良の論拠の一つを構成する。このように、2 つの尺度間の時間的進行を確実にする研究デザインは、因果関係の推論においてより説得力がある。
生物学的勾配（用量反応）	より大きなばく露またはばく露の持続時間に関連した影響の増加は、因果関係を強く示唆している。しかし、その不在は因果関係を排除するものではない。
生物学的妥当性	実験的エビデンスに基づいた生物学的に信憑性のあるメカニズムによって説明され、支持されたデータは、関連性が因果関係にある可能性を強化する。しかし、メカニズム論的データの欠如は因果関係に反するエビデンスとして捉えられるべきではない。
統一性	エビデンスの解釈は理にかなったものでなければならず、ばく露－疾病パラダイムの下で問題となっている健康影響の生物学について知られていることと矛盾するものであってはならない。もしそうであれば、ヒトに最も近い種の方がヒトとの関連性が高いと考えるべきである。
実験的エビデンス	無作為化実験の結果は、他の研究デザインに基づく結果よりも因果関係の強いエビデンスを提供する。あるいは、非実験的研究からの関連は、関連から導き出された無作為化予防が結論を導く場合には、因果関係があると考えられる。
重要事象の結果	特定の健康影響について確立された MOA/AOP の基礎となる重要事象（すなわち、in vitro、in vivo、またはヒトのデータ源を組み合わせた測定可能なパラメータ）をそれぞれ明確に説明する。完全に解明された MOA/AOP は、ヒト健康リスク評価に疫学研究を使用するための要件ではない。

Höfler（2005）、Fedak ら（2015）及び US-EPA（2016）からの引用。

予測的因果関係については、「事象 Y が事象 X の後に続いたので、事象 Y は事象 X によって引き起こされたに違いない」という論理的誤謬を避けるために注意を払わなければならない。Höfler（2005 年）は、より正確な「反事実」の定義を次のように引用している:「しかし、E があれば、D は発生しないか、発生しなかっただろうが、E があれば発生するだろう／しただろう」。しかし、記号論理を用いたより詳細な記述もある（Maldonado 及び Greenland、2002 年）。Rothman 及び Greenland（2008 年）は、「反事実効果の唯一の必須条件は、原因が効果に先行しなければならないという状態である」と述べている。結果または「影響」として提案された事象がその原因に先行している場合、事象間の関連性はあるかもしれないが、因果関係は確かにない」と述べている。

6.3.1. 疫学的証拠の統合

観察による研究のシステマティックレビューとメタアナリシスは、農薬の潜在的なハザード、ばく露反応の特徴、ばく露シナリオとばく露評価の方法、そして最終的にはリスクの特性評価の理解を強化する情報を提供することができる（van den Brandt、2002 年）。システマティックレビューは、特定のトピックに関するすべての関連研究を特定し、評価し、統合することでバイアスを低減することを目的とした、詳細で包括的な計画と事前に定められた検索戦略を伴う。システマティックレビューの主なステップは以下の通りである:研究課題の策定、包含基準と除外基準の定義、異なるデータベ

ースにまたがる研究の検索戦略、事前に定めた戦略に従った研究の選択、データ抽出とエビデンス表の作成、選択した研究の方法論的品質の評価、バイアスのリスクを含めた評価、データの統合（研究が許せばメタアナリシスを行うこともできる）、結果の解釈と結論の導き出し（EFSA、2010 年 a）。しかし、農薬疫学の分野では、標準化と調和が困難であるため、エビデンスの統合は困難である。それにもかかわらず、疫学研究の強固性と妥当性を評価する上で、エビデンス統合は極めて重要な役割を果たすべきである。

このエビデンスの評価に役立つ統計ツールが開発されている。ほぼ同一のばく露と転帰に関する複数の研究が利用可能な場合、これらの研究は重要な科学的証拠を提供することができる。ばく露と転帰が研究間で定量化され、調和されている場合には、類似した計画の個々の疫学研究からのデータを組み合わせることで、より正確なリスク推定値を得るのに十分な検出力を得ることができ、不均一性の評価を容易にすることができる。適切なシステマティックレビュー及びエビデンスの定量的な統合を定期的に行う必要がある（例えば、世界がん研究基金のがんリスク因子のメタアナリシスの継続的な更新のためのアプローチ[17]）。研究は、以前に発表された観察による研究の基準に従って評価され、可能性のある選択バイアス、測定誤差、サンプリング誤差、異質性、研究デザイン及び結果の報告と提示について慎重に検討されるべきである。

メタアナリシスとは、一般的に、異なる研究で報告された結果を組み合わせて比較するための統計的手法の集積を示すために使用される用語である（Greenland 及び O'Rourke、2008 年）。メタアナリシスの技術は、小さな研究効果や過剰な有意性バイアスなど、研究分野における多様なバイアスの存在を調べるために使用されることがある。しかし、メタアナリシスは、各研究デザインに関連している可能性のある根本的なバイアスを克服するものではない（すなわち、交絡、想起バイアス、または他のバイアスの原因が排除されない）。システマティックレビューやメタアナリシスが農薬の影響について結論を導き出すことができる範囲は、含まれた研究から得られたデータや結果が有効かどうか、つまり検討された研究の質に大きく依存する。特に、一貫したバイアスの結果として、オリジナルの研究間で一貫した結論が得られれば、システマティックレビューでは偏った結論が得られることになる。同様に、無効な研究のメタアナリシスでは、誤った影響推定値に狭い信頼性間隔が生じるなど、誤解を招く結果になる可能性がある。

レビューされた文献の基本的な研究の特徴を要約することに加えて、典型的なメタアナリシスには以下の要素が含まれるべきである。（a）対象となる各健康影響の平均影響量と影響量分布及び影響量分布の不均一性の検討（b）影響量分布に存在する変動性を系統的に解析し、効果量の大小に関連する研究の特徴を特定するサブグループ解析（c）引き出された結論の妥当性を評価するための出版バイアス解析及びその他の感度分析（Wilson 及び Tanner-Smith、2014 年）。

メタアナリシスでは、基礎となる研究集団の影響量分布を適切に記述するモデルを指定することが重要である。意味のある影響量分布を用いたメタアナリシスは、定量的リスクをリスク評価モデルに統合するのに役立つ。従来の正規の固定影響モデル及びランダム影響モデルは、パラメータと共変量に条件付きで正規の影響量母集団分布を仮定している。このようなモデルは、全体的な影響量を推定するのには適切かもしれないが、影響量分布が非正規の形状を示す場合には予測には確実に適していない（Karabatsos ら、2015 年）。

6.3.2. 研究間の異質性を探索するツールとしてのメタアナリシス

異なる研究の結果を評価する際には、多くの側面を慎重に評価すべきである。メタアナリシスを行う研究者は、調査の範囲を、考慮した研究を平均した関連性の大きさの結果に限定する傾向があります。その動機は、多くの場合、考慮した研究を平均した関連性の強さの決定に限定する傾向がある。影響の個々の推定値は偶然性によって変化するため、ある程度のばらつきは予想される。しかし、推定値は意味のある場合にのみ要約されなければならない。見落とされがちな重要な側面として、サブグループを超えた個人間の関連の強さの不均一性がある。研究間の不均一性は評価され、存在する場合には定量化される必要がある（Higgins、2008 年）。メタアナリシスでは、異なる研究からの結果の間の不均一性は、同質性と同じくらい有益であるかもしれない。観察された結果の矛盾の根底にある理由を探ること

[17] 世界がん研究基金インターナショナル。継続的更新プロジェクト（CUP） http://www.wcrf.org/int/research-we-fund/ continuous-update-project-cup

は、一般的に大きな理解につながる。

図 1 は、3 つの農薬（A、B、C）のそれぞれが 2 つの研究のメタアナリシスで評価されている仮想例の 3 つのフォレストプロットを示している。各農薬の両方の研究が、最高の品質と科学的な厳密さを持っていると仮定している。バイアスが疑われない。

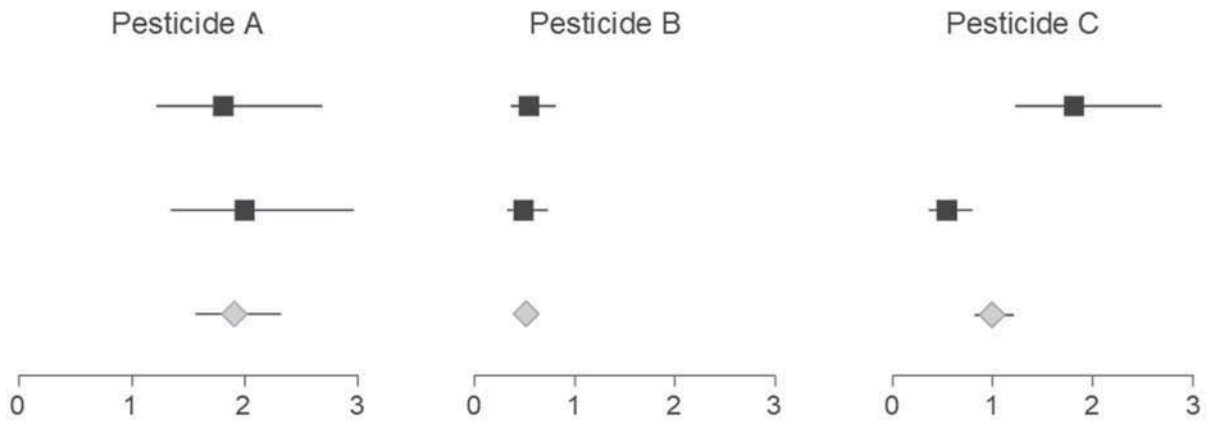


図 1:3 つの農薬（A、B、C）のそれぞれが 2 つの研究のメタアナリシスで評価されている仮想例のフォレストプロット。各プロットの x 軸は、ばく露した個体とばく露していない個体を比較した、対象疾患の推定リスク比を表している。四角は各研究における推定リスク比を示し、灰色のダイヤモンドは要約されたリスク比を示している。横線は 95%信頼区間を示す。

以下の文章には、図 1 の結果の解釈について、農薬を 1 つずつ、短くコメントしている。

- 農薬 A へのばく露は、病気のリスクを 2 倍にするようである。結果は 2 つの研究の間で一致しており、信頼区間には帰無値である 1 が含まれていない。しかし、これらの結果は、(a) 同じばく露と疾患について実施された他の研究では、リスク比が約 2 であること、または (b) 個人のどのグループ（例えば、男性か女性か、若年者か高齢者）でもリスク比が 2 であることを暗示するものではない。
- 農薬 B へのばく露は、病気のリスクを半減させるようである。結果は 2 つの研究の間で一致しており、信頼区間には帰無値である 1 は含まれていない。しかし、これらの結果は、(a) 同じばく露と疾患について実施された他の研究では、(a) リスク比が約半分になること、または (b) 個人のどのグループ（例えば、男性か女性か、若年者か高齢者か）でもリスク比が約半分になることを暗示するものではない。
- 農薬 C へのばく露は、一方の研究では病気のリスクが 2 倍になり、他方の研究ではリスクが 2 分の 1 になるようである。結果は 2 つの研究の間で矛盾しており、また、信頼区間には帰無値である 1 が含まれていない。しかし、これらの結果は、(a) 同じばく露と疾患について実施された他の研究では、リスク比が約 1 であること、または (b) 個人のどのグループ（例えば、男性か女性か、若年者か高齢者か）でもリスク比が約 1 であることを暗示するものではない。

図 1 に示された結果は、どのようなエビデンスを提供できるか？

どのような研究で報告されたリスク比も、すべての関連因子がコントロールされている場合にのみ、他の集団に一般化することができる（Bottai、2014 年；Santacatterina 及び Bottai、2015 年）。この背景では、関連因子とは、対象となる健康影響に確率的に依存する変数のことである。例えば、心血管疾患は、若年者よりも高齢者の方が多い。したがって、年齢は心血管疾患の関連因子である。図 1 に示された結果から得られるエビデンスは、検討した各研究でこのステップを踏んだ場合にのみ有効となる可能性がある。もしそうであれば、2 つの研究のそれぞれで考慮された個人の特定のグループでは、農薬 A へのばく露がリスクを 2 倍にするというエビデンスがある。リスク比がそれぞれの研究集団の要約測定値であるならば、どの結論も一般化されるべきではない。しかし、農薬 A のリスク比がいかなる因子でも調整されておらず、基礎となる母集団が 2 つの研究で大きく異なっていた場合、関連因子が存在せず、農薬 A はどのサブグ

ループの個人においてもリスクを 2 倍にするというエビデンスが残っている。農薬 B はリスクを半減させるようであり、推定された信頼区間は農薬 A よりも農薬 B の方が狭い。しかしながら、農薬 A についての上記の条件のもとでは結果を一般化できる可能性は農薬 B についても維持されている。農薬 C に関しては、フォレストプロットから、この農薬へのばく露が、一方の研究では個人のグループで病気のリスクを上げ、他方の研究ではリスクを下げるというエビデンスが得られた。繰り返しになるが、リスク比がそれぞれの研究集団の要約測定値であるならば、どの結果も一般化されるべきではない。農薬 C に関する 2 つの研究間の矛盾の背後にある理由を調査することは、農薬 A または農薬 B に関する研究間の類似性の背後にある理由を調査するのと同じくらい多くの科学的予測を提供することができる。

一般的に、図 1 の 3 つのパネルのそれぞれにある銀色のダイヤモンドのようなフォレストプロットによって提供される全体的な要約評価尺度は、ほとんど科学的な関係を持たない。異なる研究の結果を評価する際には、多くの側面を慎重に評価しなければならない。見落とされがちな重要な側面は、サブグループを超えた個人間の関連の強さの不均一性である。研究を記述した出版物でサブグループ解析に関する情報が提供されている場合、これは慎重に評価されるべきである。感度分析は、異なる研究で得られた結果を補完すべきである。これらの解析は、異質性と、情報とサンプリング誤差とともに、関連する因子を制御していない場合の影響を評価することを目的とすべきである。図 2 に総観図を示す。

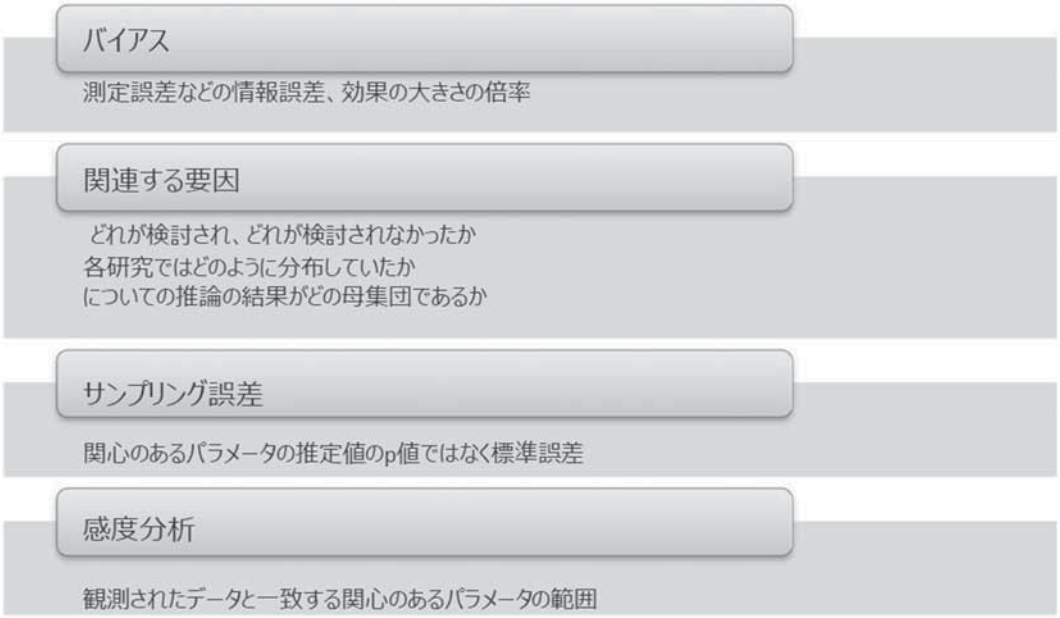


図 2:複数の研究を評価・比較する際に考慮すべき項目

6.3.3. ハザード同定のためのメタアナリシスの有用性

ヒトのデータはリスク評価の多くの段階で利用できる。単一の疫学研究(同じ農薬に関する追加研究が入手できない場合)は、質の高い研究(表 2 に示す基準による)でない限り、単独のハザード同定のための情報源として使用すべきではない。代わりに、システマティックレビューやメタアナリシス(必要に応じて)など、多くの研究をまとめたエビデンス統合技術を利用すべきである。慢性疾患に関連するデータの定量的な統合のために多くのメタアナリシスが実施されているが、リスク評価モデリングへの応用はまだ限られている。

重要なことは、エビデンス統合は、現在のエビデンスの方法論的評価とバイアスのリスク評価を提供し、不確実性の領域を強調し、強固で信頼性の高いエビデンスとの関連を特定することである。

図 3 は、疫学研究をリスク評価に適用するために提案された簡単な方法論を示している。最初の考慮事項は、同じ健康影響を扱う異なる疫学研究を組み合わせる必要性である。これは、EFSA のシステマティックレビューのためのガイダンス(EFSA、2010 年 a)で提案されている基準に従って行うことができる。次に、研究デザインと実施、母集団、ばく露評価、健康影響評価、交絡因子の管理、統計解析、結果の報告など、WOE 評価のための 6.2 節に記載されてい

る要素に基づいてバイアスのリスクを評価する。信頼性が低いと分類された研究は、リスク評価のために受け入れられないと考えられる。残りの研究は重み付けを行い、ハザードの特定に使用する。

定量的データが利用可能な場合、メタアナリシスを実施して要約データを作成し、利用可能な、あるいは選択基準を満たすすべての個々の研究の結果を組み合わせることで、統計的な検出力とリスク推定値（OR、RR）の精度を向上させることができる。メタアナリシスは、関連の大きさが検討した研究の平均値に決定するので、ハザード同定のためのより強力な基盤を提供する。さらに、特定の状況下では、健康影響におけるこれらの測定された差（OR、RR）を用量反応関係に変換できるため、リスク特性の測定基準に移行する可能性がある（Nachman ら、2011 年）。実際には非常に珍しいことではあるが、これにより、動物からの外挿法を使用することなく、ヒトにおける重大影響の同定や基準値の設定が可能になる。

メタアナリシスでは不均一性が一般的であるため、どの研究を定量的に組み合わせることができるかを評価する必要がある。異質性は、異なるサブグループにおける多様な影響を表す真正なものである場合もあれば、バイアスの存在を表す場合もある。異質性が高い場合（I^2 が 50%を超える場合）は、異なる情報源からのバイアスを集約するリスクが高いため、要約尺度を得るために個々の研究を組み合わせるべきではない。感度分析及び／またはメタ回帰によって異質性の原因を探るべきである。さらに、メタアナリシスにおける多様なバイアスの存在、例えば、小規模な研究効果、出版バイアス、過剰な有意性バイアスなどを調べるべきである。基礎となる研究集団の影響量分布を適切に記述するモデルをみつけることが重要である。

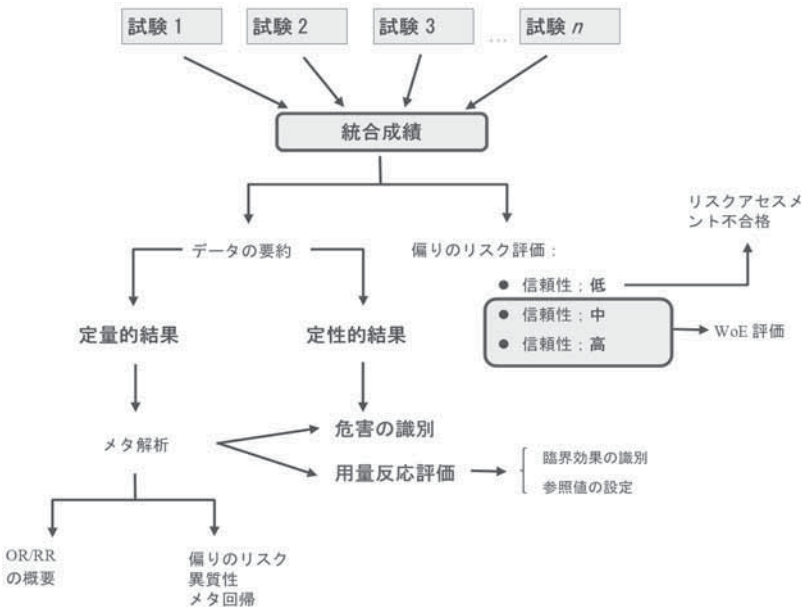


図3：疫学研究をリスク評価に活用するための方法論

6.3.4. 潜在的な用量反応モデル化のための類似の疫学研究からのデータの蓄積

他の研究と同様に、単一の疫学研究から得られた結果は、複製によって検証する価値がある。複製の数が豊富な場合には、メタアナリシスによって複製された疫学研究の全セットを評価し、主要な健康影響事象について、研究間で結果が一貫しているかどうかを確認することについては価値があるかもしれない。このようなアプローチにより、因果関係の存在についてより強固な結論が得られるであろう。

ハザードが同定されると、リスク評価の次のステップは、異なるばく露レベルでの有害な影響のリスクを推定するための用量反応評価を実施することであり、特定の集団に対して健康への有害な影響が認められない濃度以下のばく露レベルを推定する。しかし、このステップでは、個人レベルでの完全に定量的な（または少なくとも半定量的な）ばく露データが必要である。定量的統合から得られた要約推定値は、研究間の相対比較が可能となるようなばく露の連続変数の変化（またはばく露のある一定割合の変化）に対する OR を示すものであれば、リスク評価にとってより有益であり、研

究間の相対比較が可能となり、健康に基づく基準値を導き出すのに役立つ。このような枠組みの中でのみ、類似した計画のヒト研究からのデータを統合して、適切な用量反応曲線をモデル化するのに十分な力を得ることができる（Greenland 及び Longnecker、1992 年;Orsini ら、2012 年）。

逆に、メタアナリシス的アプローチは、すべての研究で同じ割合で含まれる有効成分へのばく露が必ずしも必要ではないため、ばく露を「はい」または「いいえ」（これまでにあり、または決してない）と解釈するメタアナリシスに基づいて複合 OR が計算された場合には、限られた価値しかないかもしれない。これらのケースでは、メタアナリシスは一貫して農薬ばく露に関連したリスクの増加を示しているが、リスク評価のためには、ばく露は特定の農薬クラスの影響を特性評価する必要があり、同じクラス内でも効力が異なる可能性があるため、個々の農薬の影響を特性評価する必要がある（Hernándezら、2016 年）。

このアプローチは、Point of Departure を特定することを可能にし（例えば、ベンチマーク用量（BMD））、疫学研究を定量的リスク評価に統合することに関連するであろう。現在、BMD モデリングは実験研究からの用量反応データの解析に使用されているが、観察による疫学研究からのデータにも同様のアプローチを適用することが可能である（Budtz-Jørgenson ら、2004 年）。EFSA 科学委員会は、BMD アプローチは、実験研究や疫学研究からの用量反応データを利用して潜在的なリスクをよりよく特徴付け、定量化するために、Reference Point を得るための NOAEL（no observed-adverse-effect level）アプローチと比較して、より科学的に進んだ方法であると結論づけている。このアプローチは、原則としてヒトのデータにも適用可能である（EFSA Scientific Committee、2017 年 b）が、対応するガイドラインはまだ作成されていない。

観察による疫学研究からの用量反応データは、いくつかの点で典型的な動物試験の毒性データとは異なる可能性があり、これらの違いは BMD の計算に関連する。ばく露データは、多くの場合、少数の十分に定義された用量群に当てはまらないことが多い。ほとんどの実験研究とは異なり、観察による研究には完全に未ばく露の対照群が含まれていない場合がある。この場合、用量反応曲線を作成することは必ずしもばく露量ゼロでの観察を必要としないため、BMD アプローチが適用される。しかし、ばく露量ゼロでの反応は低用量外挿法で推定する必要がある。したがって、疫学データから得られる BMD はモデル依存性が強い（Budtz-Jørgensen ら、2001 年）。

疫学データは、BMD アプローチを適用するためには、特に特定の農薬とそのばく露に影響を与えるという点で、十分な品質のものでなければならない。このような BMD アプローチについては、明確なルールとガイダンス、モデルパラメータの定義を考慮する必要があり、制御された実験環境からの BMD アプローチとは異なる可能性がある。BMD モデリングアプローチは重金属やアルコールに関する疫学データに適用されているが（Lachenmeier ら、2011 年）、現在のところ、農薬に関する個別の研究は、用量反応モデリングに使用するのに適しているものはほとんどなく、他の研究と組み合わせて使用することはあまりない。しかし、今後も研究は実施され、同様の報告がなされ、それらの研究を蓄積して、より強固な評価を行うことができるようにすべきである。

7. 多様なエビデンスの統合：ヒト（疫学データと警戒データ）と実験の情報

本節では、まず第 7.1 節で、実験研究や疫学研究に由来する主なエビデンスの特性の違いについて考察する。使用したアプローチは、EFSA Scientific Committee Guidance on WOE（EFSA Scientific Committee、2017 年 b）で推奨されているもので、これらの異なる情報を評価し統合するために、信頼性、関連性、一貫性の 3 つの段階を区別している。最初の段階では、疫学的研究（セクション 6 で述べている）や実験的研究（この意見書の範囲を超えている）である個々の研究の信頼性の評価で構成されている。次に、信頼性があると判断された 1 件以上の研究の関連性（エビデンスの確実性）を疫学（第 6 節に記載）と毒物学の原則を用いて評価する。次に、第 7.2 節では、WOE アプローチで検討された疫学的及び実験的研究からの様々な関連情報を、ヒトに対する一貫性と生物学的妥当性を評価するために、どのようにしてまとめるかを検討する。

7.1. 異なるエビデンスの起源と特性 実験的アプローチと疫学的アプローチの比較

農薬の規制リスク評価では、毒性影響に関する情報は、規則(EC)283/2013 及び 284/2013 で要求され、OECD ガイドラインに従って実施された一連の実験結果に基づいている。これらの実験は in vivo または in vitro で実施されているため、規則(EC)1107/2009 に基づき申請者が提供することが要求されているように、農薬については常に質の高い実験データが存在している。EFSA の有効成分のピアレビューによると、各エビデンスの信頼性を評価するために、許容範囲、補足範囲、不許容範囲といういくつかのカテゴリーが設定されている。in vivo または in vitro 毒性試験のデータの質及び信頼性は、ハザード及びリスク評価のための試験の妥当性を判断するためのより構造的な裏付けをより良く提供する評価方法を用いて評価されるべきである。農薬の健康リスク評価に使用できるように、実験的研究の実施と報告のための基準が提案されている(Kaltenhäuser ら、2017 年)。

標準化された試験ガイドラインや優良試験所規範(GLP、例えば OECD 試験ガイドライン)に従って実施された農薬有効成分の動物(in vivo)試験は、通常、他の研究よりも信頼性が高いとされている。しかし、このような枠組みの下で実施された研究の方がバイアスのリスクが低いというエビデンスはないため(Vandenberg ら、2016 年)、GLP と非 GLP の両方を問わず、関連するすべての研究から得られたエビデンスも考慮し、重み付けを行うべきである。このように、査読された科学文献からのデータは、方法論的信頼性を評価した後に十分な品質のものであれば、農薬有効成分の規制リスク評価のために考慮されるべきである。WOE 全体への貢献は、試験系、試験計画、統計的方法、試験項目の特定、結果の文書化、報告などの要因によって左右される(Kaltenhäuser ら、2017 年)。

ハザード及びリスク評価のための試験の妥当性を判断するためのより良い補助とするために、in vitro 毒性試験の内部妥当性も評価されるべきである。in silico モデルは、構造活性相関(SAR)を導出し、ヒトにおける有効成分の作用様式や作用機序の同定や特性評価のための現行の毒性試験を補完するために使用することができる。これらの代替毒性試験(及び非試験)アプローチは、動物データがない場合、例えば農薬の潜在的な神経発達や内分泌かく乱作用をスクリーニングし、動物試験の信頼性を高めるのに役立つ可能性がある。規制目的のために動物試験の数を最小限にすることが求められていることを考慮すると、動物試験以外の情報は WOE 評価に使用できる適切な独立したエビデンスを提供することができる。

化学物質のヒト健康への影響から特定の化学物質へのばく露に伴うリスク、化学物質の混合物の毒性、毒性反応のバイオマーカーの関連性、あるいは新しい毒性試験法の評価まで、多くの毒性学的問題がシステマティックレビューの対象となっている(Hoffmann ら、2017 年)。例えば、以前の Scientific Opinion では、EFSA は AOP アプローチの枠内で毒性学的メカニズムの決定にシステマティックレビューを使用した(Choi ら、2016 年;EFSA Scientific Committee、2017 年 c)。

有効成分の毒性データの他に、食事や飲料水を通じてヒトにばく露された場合には、代謝物や残留物についてもデータが必要となる場合がある。これらの研究から得られた結果は、食品消費やその他のばく露源から推定される予想されるヒトばく露量との関連で検討される。このアプローチの長所は、動物の代謝経路が必ずしもヒトの代謝経路と類似しているとは限らないにもかかわらず、in vivo 試験では潜在的な毒性代謝物を考慮していることである。

実験動物を用いた実験研究は、交絡因子が排除されるように計画された研究であるが、疫学研究では必ずしもそうとは限らない。しかし、規制研究に使用される動物は、一般的には近親交配されて遺伝的に同一であり、制御された環境のために、定量的及び定性的な化学物質感受性のすべての特性を欠いている。それにもかかわらず、ヒト疾患の動物代替は、その科学的妥当性とヒトへの外挿性が問われており、動物データとヒトの結果との間にしばしば見られる相関性の欠如は、疾患経路や遺伝子発現プロフィルにおける疾患誘発性の変化における実質的な種間差に起因していると考えられている(Esch ら、2015 年)。そのため、多くの実験モデルは複雑な多因子疾患を捉えていないため、動物からヒトへの外挿はかなりの不確実性を伴うものとなっている。したがって、現在のリスク評価はその特性上、予測的なものであり、化学物質に特化したものであり、ヒトは環境、食事、職業上の源からの多数の化学物質にばく露されているため、あるいは異なる毒物動態の違いのために、十分ではないかもしれない。動物からヒトへの外挿の不確実性を考慮して、規制当局のリスク評価のアドバイスは、単に安全性が確認されている関連の Point of Departure(NOAEL、

LOAEL、または BMDL）を考慮するのではなく、不確実性因子（UF）を用いてこれらの値を下げ、急性毒性または慢性毒性の安全な参照用量値を提案するものである。

実験動物を用いた研究の限界を考えると、たとえそれ自体に限界があるとしても、「実社会」での疫学研究が必要である。疫学的研究では、集団のばく露の真の範囲（または推定された範囲）を組み入れているが、これは通常、一貫した速度と用量の大きさで発生するのではなく、断続的で一貫性のない用量で発生する（Nachman ら、2011 年）。疫学的研究は実社会のばく露に基づいているため、実際のヒトのばく露についての予測を提供し、それを疾病に結びつけることができ、種を超えた外挿に関連する不確実性を回避することができる。したがって、リスク評価は、優れた植物保護対策（Good Plant Protection Practice）と現実的な使用条件に基づいて行われるべきであると規定した規制 1107/2009 の第 4 条の要件に対応していると言える。このように、疫学的研究は、高用量外挿の必要性を回避しつつ、問題の定式化とハザード／リスクの特性評価を支援する（US-EPA、2010 年）。

したがって、疫学的研究は、(a) 動物モデルでは検出が困難な特定のヒトの健康影響との関連性を特定する。(b) 動物モデルで特定された影響のヒトへの関連性を明らかにする (c) 動物モデルが利用できない、または限定された、健康影響を評価する能力を提供する（Raffaele ら、2011 年）。疫学的証拠は、十分に強固な農薬の疫学的研究が利用可能な場合にのみ、動物実験によるエビデンスよりも考慮される。しかし、疫学研究では、健康影響に影響を与え、結果を混乱させる様々な要因が常に存在する。例えば、疫学的データが農薬製剤へのばく露が有害であることを示唆していても、通常、農薬へのヒトばく露を正確に評価することの複雑さから、どの成分が原因であるかを特定することはできない。一部の製剤補助剤は本質的には毒性がないが、有効成分の毒物動態を変化させる場合には、毒性学的に関連性がある可能性がある。さらに、ばく露に関連した測定されていない因子による交絡を完全に排除することはできないが、仮説的な交絡因子（まだ認識されていない）は実際の交絡因子ではない可能性があり、リスク（または効果量）の推定値に意味のある影響を与えるためには、疾患やばく露と強く関連していなければならないが、必ずしもそうとは限らない。

多くの病気は複数のリスク因子と関連していることが知られているが、脆弱なシステムに対する個々の農薬のハザードの影響を評価するためには、通常、ハザードごとのアプローチが考慮される（図 4A）。特に、単一リスク解析では、特定の条件下で発生する特定のハザードやプロセスに起因する個々のリスクを決定することができるが、異なる環境ストレス要因（自然現象または人為的要因のいずれか）によって引き起こされる複数のリスクを統合的に評価することはできない（図 4B）。リスク評価は、複数の有害な健康影響の同時発生のエビデンスを評価するための手順を開発することが有用である（Nachman ら、2011 年）が、これはよりヒトの環境で起こることと一致している。これらの理由から、適切に実施されれば、疫学研究はリスク評価プロセスとの関連性が高い。

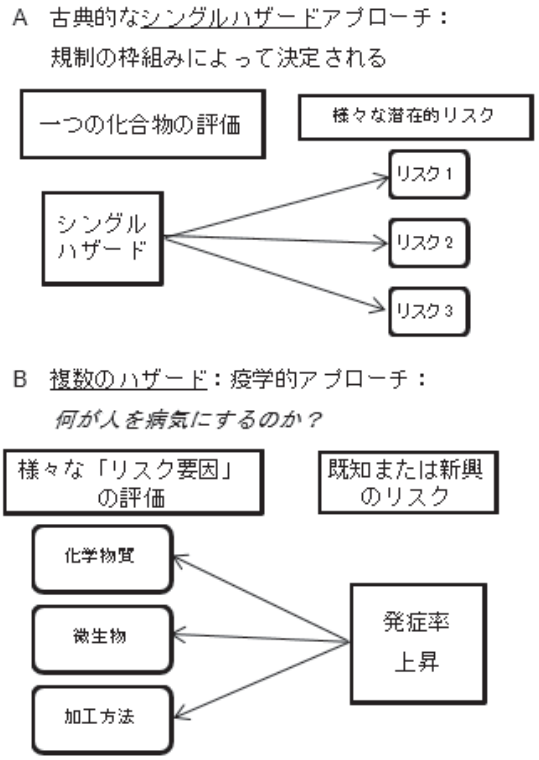


図 4: 古典的な毒性学的研究と比較した場合の疫学的研究の役割

疫学的データと並行して、特に急性毒性については、警戒データが追加のエビデンスの流れを提供することができる。通常、症例は十分に文書化されており、リスク評価のさまざまな段階で情報を利用することができる。これらの情報には、ばく露のレベルと期間、臨床経過、因果関係の評価などが含まれる。重度のケースでは、毒素及び／または代謝物が通常、血液または尿中で測定され、動物データとの比較が可能であり、場合によっては毒性値を設定することができる。

要約すると、実験的研究、疫学的研究、及び警戒データは、エビデンスを収集し評価するための 2 つの異なるアプローチを表している。すなわち、実験的な研究計画と比較的均質な代理母集団を用いた（通常は単一物質に対する）管理されたばく露から得られるものと、非実験的な研究デザインを用いた混合ばく露条件（及び変化する）から不均質な対象集団で観察される変化を反映したものである（ECETOC、2009 年）。疫学と毒物学は、それぞれがヒトへのハザードの特定に重要かつ異なる貢献をしている。このことは、両方のエビデンスの流れを補完的なものにしており、それらを組み合わせることで強力なアプローチが可能になる。動物実験は常に疫学的研究の解釈に情報を与えるものであるべきであり、その逆もまた然りであるため、これらを独立して研究・解釈すべきではない。

7.2. ヒトの観察データと実験動物の実験データの重み付けの原則

信頼性の高いヒト（疫学的または警戒研究）研究の特定し、プールされたヒト研究の関連性の評価した後、関連性があると判断された別個のエビデンスを、同様に関連性があると判断された他のエビデンスと統合する必要がある。

したがって、最初の検討事項は、検討対象の健康影響がどれだけ毒性学的及び疫学的研究でカバーされているかということである。特定の健康影響／エンドポイントについて動物試験とヒト試験の両方が利用可能であると考えられる場合、これは、様々なエビデンス源の重み付けに先立って、個々の試験が信頼性とエビデンスの強固さを評価されることになる（疫学的研究については、それぞれ 6.2 項と 6.3 項）。異なるデータセットは補完的で確実なものであるが、個別には不十分な場合があり、ヒトの健康リスクを適切に特定するための課題となる可能性がある。良好な観察データが

不足している場合は、実験データを使用しなければならない。逆に、実験データが利用できない場合や、既存の実験データがヒトに関連していないことが判明した場合には、リスク評価は利用可能で適切な観察による研究に頼らなければならないかもしれない。

リスク評価のために、複数のエビデンス(特にヒト研究と実験研究)から得られたデータを体系的に統合するためのフレームワークが提案されている(図 5)。このような統合は、修正された Bradford Hill 基準(表 3)を用いて、関連性、一貫性、生物学的妥当性を考慮した WOE 解析に基づいている。ヒトと動物のデータを比較解釈するためには、この枠組みは以下の原則に依存すべきである(ECETOC、2009 年; Lavelle ら、2012 年)。

- エビデンスの全体を評価すべきであるが、信頼性があると判断された研究(許容可能なエビデンスまたは補足的エビデンスに分類された研究)のみがさらに検討される。ヒト研究または実験研究から得られたデータの信頼性が低い(許容できないと分類される)と考えられる場合は、リスク評価を行うことはできない。
- 複数のエビデンスが関連していることが判明した場合には、WOE アプローチに従うべきである。農薬有効成分については、OECD の試験ガイドラインに従った実験研究は、それに反するエビデンスがない限り、信頼性が高いとみなされる。動物実験からのエビデンスの強固さは、代替的な農薬毒性試験または非試験方法(例えば、それぞれ in vitro 試験と in silico 試験)に高い信頼性がある場合に向上させることができる。疫学的証拠については、メタアナリシスを実施することで、個々の研究よりも効果の大きさをより正確に推定でき、また、研究間のばらつきを調べることができる(第 6.3 節参照)。
- 次に、評価されるステージに関連性が高いと判断された研究は、データがヒトまたは動物の研究からのものであるかどうかに関わらず、より重要視されるべきである。ヒトのデータが最も関連性が高く、作用機序的な科学的根拠に支えられている場合には、リスク評価の各段階で優先されるべきである。ヒトデータと実験データの関連性が同等または類似している場合には、どのデータセットが優先されるかを判断するために、それらのデータの一致性(エビデンス系統間の一貫性)を評価することが重要である。
  - ヒトデータと実験データが一致している場合、リスク評価では、ハザードの特定(例:両方とも同じハザードを示す)またはハザードの特性評価(例:両方とも同じような安全用量を示唆する)のいずれにおいても類似した結果が得られるため、すべてのデータを使用すべきである。このように、両者は互いに補強し合うことができ、両方のケースで同様のメカニズムを想定することができる。
  - 不一致の場合、フレームワークはこの不確実性を考慮する必要がある。ハザードの特定については、一般的にハザードの存在を示唆するデータが優先されるべきである。用量反応については、低い許容量のデータが優先されるべきである。不一致が生じた場合には、この相違の理由を検討すべきである。その理由が基礎となる生物学的メカニズム、またはヒトと動物モデル間の毒物動態学的差異に関連している場合は、リスク評価の信頼性が高まる。逆に、その理由が理解できない、あるいは説明できない場合は、リスク評価の信頼性が低下するかもしれない。このような場合には、生物学的エビデンスにおける矛盾に対する考察を行うべきである。

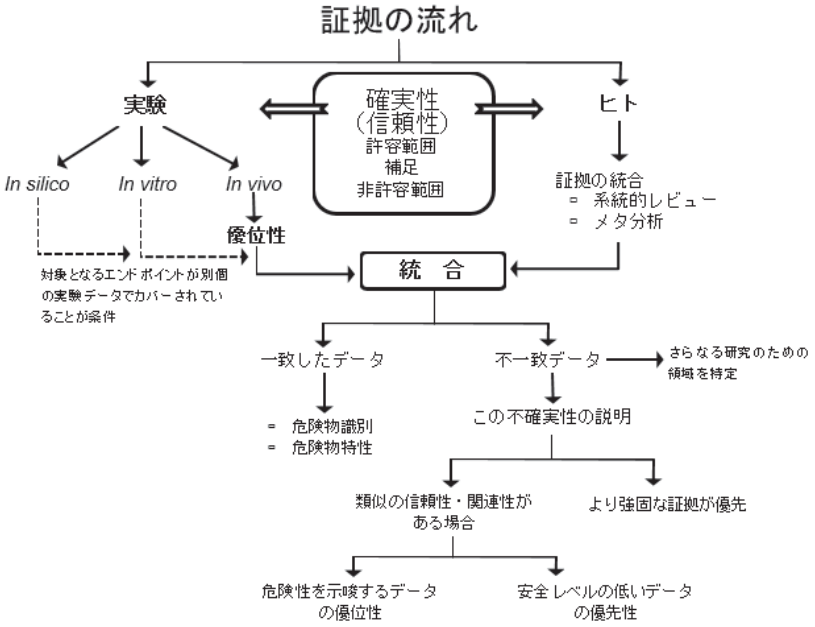


図 5:リスク評価のためのヒトと動物のデータ統合の方法論

疫学研究はリスクを解析するための補完的なデータを提供するものであり、十分に計画された毒物学的 in vivo 研究やメカニズム研究と併せて文脈を考慮する必要がある。複数のエビデンスを統合して得られたエビデンスの総合的な強度は、少なくとも単一のエビデンスで得られた最高のエビデンスと同程度になる。この統合的アプローチは、毒性学的証拠と疫学的証拠をどのように重み付けして統合するかについての明確な指針を提供するものである。これは複雑な作業であり、疫学的データが多因子性、多発性、慢性の疾患を対象としている場合には、毒物学的モデルや疾患特異的動物モデルが限られているため、さらに困難な作業となる。

7.3. 異なる起源のエビデンスの重み付け

WHO/IPCS は、WOE アプローチをリスク評価に関連すると考えられるすべてのエビデンスが評価され、重み付けされるプロセスと定義している(WHO/IPCS、2009 年)。WOE アプローチは、化学物質のリスク評価を例にとると、異なる一連のエビデンス(in vivo、in vitro、in silico、母集団研究、モデル化されたばく露データや測定されたばく露データなど)の評価を必要とする。課題は、体系的で一貫性がある明確な方法で、これらのタイプのエビデンスを重み付けすることである(SCENIHR、2012 年)。重み付けは、形式的には定量的なものであってもよいし、リスクの基準を参照した分類に依存するものであってもよい。

EFSA のパネルと科学技術委員会による科学技術データの評価に WOE アプローチを使用するための明確な基準を提供するために EFSA ワーキンググループが設立された(EFSA、2015 年 b)。このワーキンググループの目的は、個々の研究がどのように選択され、どのように重み付けされるべきか、結論に到達するためにどのように知見を統合し、結論に関する不確実性をどのように特定するかについて、利害関係者に支援を提供することであった。

WOE アプローチは、DAR や RAR のピアレビュープロセスにおける農薬のリスク評価において一貫して考慮されていない。構造化された WOE アプローチを用いずに専門家の判断だけで行うことがより一般的になっている。いくつかの例を挙げると、グリホサートのピアレビュー(EFSA、2015 年 c)では、報告者である加盟国(RMS)は、疫学的データを含む産業界または公的文献からのすべてのデータを考慮し、確立された事後基準と、探索された毒性の各エンドポイントについて「全体的な」NOAEL を提案するために利用可能なすべてのデータを考慮した特定の WOE アプローチを採用している。

US-EPA は最近、「健康リスク評価にヒトの疫学的データと事象データを組み込むためのフレームワーク」に従って、

農薬のクロルピリホスのピアレビューに WOE アプローチのための特定の基準を適用した。この特定のケースでは、実験的毒性研究、疫学研究、生理学的ベースの薬物動態・薬力学（PBPK-PD）モデリングを含む多くのエビデンスを横断して、定量的・定性的な知見を統合するために WOE 解析が実施された。クロルピリホスは、規則（EC）No 1107/2009 に基づく文献検索に関する EFSA ガイダンスの例としても使用されている。さらに、EFSA の結論（EFSA、2014 年 a）は、2011 年に出された最初の結論を修正するために US-EPA のレビュー（2011 年）を考慮に入れている。

まとめると、根本的な因果関係の可能性を高めるための組織的なツールとして、修正された Bradford Hill 基準を使用して、利用可能な科学的データを評価するために、より広範な WOE アプローチを適用することができる（表 3）。疫学は因果関係の立証にますます貢献しているが、この目的のための重要なステップは、生物学的妥当性の立証である（US-EPA、2010 年；Adami ら、2011；Buonsante ら、2014 年）。

7.4. 健康影響の基礎となる生物学的メカニズム

生物学的メカニズムとは、農薬とその生物学的標的との相互作用に続く健康影響につながる主要なステップを記述したものである。毒性のメカニズムは、健康への悪影響につながる主要なステップとして記述されている。影響につながるすべてのステップを理解する必要はないが、化学的相互作用に続く重要なイベントを特定することは、メカニズム（健康に悪影響を及ぼす場合の毒性）を記述するために必要である。多くの疫学研究で農薬ばく露と慢性疾患との関連性が示されているが、ヒトの疫学的観察にメカニズム的な裏付けと生物学的な妥当性を与えるためには、補完的な実験研究が必要である。実験でのばく露は、実験動物の生物学的メカニズムがヒトで発生することを条件に、ヒトの集団に関連するものでなければならない。

疫学研究の解釈の一部として生物学的妥当性を確立することは重要であり、最新の技術とアプローチを活用すべきである（7.6 節）。この意味では、AOP の枠組みは、毒性結果の基礎となる生物学的メカニズムを調査し、実験研究と観察による研究の両方で観察された関連性の因果関係を知るために、異なる情報源からの複雑な情報を体系的に整理し、統合するためのツールとして利用できる（7.5 節）。

農薬の潜在的な毒性作用の基礎となる生物学的メカニズムや経路について特定の情報を提供するためのデータの利用は、特定の健康影響に関連した生物学的機能について調査された農薬化学物質のみであるため、限られたものである。特に、同等の動物実験の結果の間に一致点がある場合や、異なる化学物質が同じ毒性パターンを示す場合には、作用機序（MOA）仮説を立てることが可能な場合がある。毒物と標的臓器を特定することは、考えられる影響の用量反応曲線とその時間的関係と同様に不可欠である。毒性につながるさまざまな主要事象と MOA 仮説を特定できれば、ヒトに対するこれらの事象の妥当性を評価することが可能になることもある（ECETOC、2009 年）。

農薬のスルホキサフロル（Sulfoxaflor）は、MOA が広範囲に研究されてきた例であり、2014 年 11 月に開催された ECHA/EFSA MOA/HRF ワークショップでも広く取り上げられている。スルホキサフロルはラットとマウスの両方で肝発がん性を誘発した。これらの肝腫瘍に対する MOA を決定するための試験は、発がん性試験が終了する前または終了するまでに MOA データが入手できるように、標準的な毒性試験のバッテリーの一部として、統合的かつ前向きな方法で実施された。WOE アプローチで評価された MOA データは、スルホキサフロルの齧歯類での肝腫瘍の MOA がヒトでは発生しないことを示している。この理由から、スルホキサフロルはヒトの潜在的な肝臓発がん性物質ではないと考えられている。

さらに、MOA データが影響の可能性がないことを示す場合もある。仮に、ヒトでは有害な影響が発生する可能性がないことを示す生物学的データがある場合は、疫学研究の解釈に反映させる必要がある。それにもかかわらず、害虫とヒトの間の一次標的部位選択性は農薬の安全性において重要な役割を果たしているが、哺乳類における二次標的も考慮しなければならない。

複数の農薬にばく露した場合、リスクを組み合わせるかどうかの判断は、農薬が共通の毒性メカニズム（同じ標的組織で同じ分子標的に作用し、同じ生化学的作用機序で作用し、共通の毒性中間体を共有する）を共有しており、同じ重大影響を引き起こす可能性がある場合、あるいは単に同じ標的臓器を共有しているという観察に基づいて判断することができる（EFSA、2013 年 a,b）。しかし、累積リスク評価は本意見書の範囲を超えている。

7.5. 有害転帰経路（AOPs）

AOP の解析法は、リスク評価に有用な方法で、関連する化学的、生物学的、毒物学的な情報を収集し、評価するための枠組みを提供している（OECD、2013 年）。AOP は、化学物質と生物学的標的との相互作用（標的分子への作用（Molecular initiating event：MIE））から、ヒトの健康に関連する in vivo での有害な事象に至るまでの一連の重要事象として定義することができる。これらの重要事象はすべて MOA の必要な要素であり、経験的に観察可能であるか、またはそのような事象の生物学的ベースのマーカーを構成するものでなければならない。したがって、AOP は、リスク評価に関連する生物学的組織のレベルで、1 つの MIE から 1 つの有害な事象に至る直線的な経路である。AOP の目的は、因果関係の連鎖において、1 つの MIE から 1 つの有害な事象へとつながる重要事象の段階を記述するための、柔軟なフレームワークを提供することである（EFSA PPR パネル、2017 年）。重要な事象は実験的に測定可能でなければならず、有害な事象は通常、in vivo での OECD 試験ガイドラインに関連している。しかし、場合によっては、有害な事象が試験ガイドラインに記載されている先端エンドポイントよりも低いレベルの生物学的組織のレベルにあることもある（OECD、2013 年）。

特定の MIE はいくつかの毒性影響につながる可能性があり、逆に、複数の MIE は同じ毒性影響に収束する可能性がある。しかし、各 AOP は 1 つの MIE と 1 つの毒性影響象のみであるが、無制限の数の中間ステップを伴うこともある（Vinken、2013 年）。生物学的組織の異なるレベルでの重要事象は、同じレベルの組織での複数の事象よりも大きな WOE をもたらすことに留意すべきである（OECD、2013 年）。

毒性反応に関与する重要な生化学的ステップは、関連する科学文献の詳細な調査や実験研究から同定され、検索される。構造データ、「オミクスベース」データ、in vitro、in vivo、あるいは in silico のデータなど、あらゆるタイプの情報を AOP に組み込むことができる。しかし、in vitro データよりも in vivo データの方が優先され、対象となるエンドポイントは代替エンドポイントよりも優先される（Vinken、2013 年）。特定された AOP は、生物学的に妥当性をもつ必要があるため、正常な生物学的プロセスと両立しないものであってはならない。

定性的 AOP（AOP 開発のための OECD ガイダンスに従った WOE の組み立てと評価を含む AOP として意図されている）は、経路に悪影響を与える農薬へのばく露と有害な影響との間の関連性の生物学的妥当性を裏付ける（または裏付けがないことを特定する）ことによって、疫学研究をリスク評価に統合するプロセスの出発点であり、標準的なアプローチであるべきである。したがって、定性的な AOP は、メカニズム的知見に基づく疫学研究の生物学的妥当性を裏付けるために、ハザード同定の目的のためだけに開発される可能性がある（EFSA PPR パネル、2017 年）。

AOP フレームワークは、異なる情報源から収集された複雑な情報のレビュー、整理、解釈を行うための柔軟で透明性の高いツールである。このアプローチには、因果関係の推論に関連する不確実性を定性的に特徴づけ、不確実性を低減するためには、追加のメカニズム的研究や疫学研究がより効果的であるかどうかを特定するという付加的な利点がある。したがって、AOP フレームワークは、有害な影響が生物学的にもっともらしいかどうかを探るためのリスク評価に有用なツールである。生物学的に妥当性を解析する目的では、AOP は重要なツールとなり得る。特に、規制上の動物を用いた毒性試験が陰性であっても、疫学研究で観察された先端エンドポイント（または関連するバイオマーカー）の評価が AOP に基づいて不十分であると考えられる場合には、AOP は重要なツールとなり得る。先端エンドポイントをメカニズム的に記述することにより、AOP はリスク評価におけるハザードの特定と特性評価のステップに貢献する。AOP フレームワークは化学的には不明確な点があるため、MOA 及び／または試験・評価に関する統合的アプローチ（IATA）フレームワークで補完されれば化学物質特有のリスク評価をサポートすることになる（EFSA PPR パネル、2017 年）。

AOP と MOA のデータは、疫学研究の結果を評価し、その結論に重み付けをするために使用できる。それらの結論が生物学的メカニズムの深い理解と矛盾するものであろうと、あるいは単に経験的なものであろうと、一度確立された AOP や MOA の枠組みと一致する他の結論よりも、それらの結論の重要性は低く設定されるべきである。しかし、十分に文書化された AOP の例は比較的少なく、完全な AOP/MOA の枠組みはリスク評価に疫学的研究を用いるための要件ではない。

したがって、AOP は、動物実験で観察された末端の影響（apical effects）に大きく依存する現在の試験パラダイム

ではなく、メカニズムに基づくリスク評価への移行を促進するための重要な要素である。リスク評価のパラダイムをメカニズム的な理解へと移行させることで、単一の農薬のヒトへの健康影響を予測する上での動物データの限界を下げることになり、また、農薬ばく露の累積リスク評価に関する現在の取り組みを支援することにもなる（EFSA PPR パネル、2017 年）。

7.6. 毒性の基礎となる生物学的経路とメカニズムを特定するための新しいツール

毒性経路の解明は、特にバイオモニタリング、オミクス技術、システムバイオロジー（毒性学）の進歩から、顕在化した疾患への毒性力学的進行における初期の生物学的摂動の新規バイオマーカーを特定する機会をもたらしている。疫学におけるオミクスの革命は、早期効果の新しいバイオマーカーの可能性を秘めており、関連のメカニズム、生化学的経路、因果関係を調査する機会を提供している。

疫学調査におけるバイオモニタリングデータの価値が認識されつつあることは、ばく露と転帰の客観的な尺度を提供することで、誤分類を減らすのに役立つかもしれない。ばく露、転帰、感受性に関するバイオマーカーデータが増加している限り、疫学は農薬ばく露の関数としての毒性力学的進行の理解と最終的にはリスク評価に大きな影響を与えることになるであろう。リスク評価者にとっての課題は、毒性力学的経路に沿った軽微で初期の変化が、下流への影響の可能性の増大を示唆していることを認識することである（Nachman ら、2011 年）。オミクスデータは、農薬の影響を受ける経路を特定することで MOA への理解を深め、リスク評価の第一段階であるハザードの特定を支援することができる。

生物学的サンプルのトランスクリプトーム（Transcriptomic,）、メタボローム（metabolomic）、エピゲノム（epigenomic）、プロテオミクス（proteomic）のプロファイリングは、環境化学物質の影響下での細胞の進化状態の詳細な画像を、時には個々の分子レベルで提供し潜在的な健康影響との早期のメカニズム的な関連性を明らかにすることができる。今日では、オミクス技術の進歩が規制毒性学にもたらす課題と長所については、まだ研究が進められている（Marx-Stoelting ら、2015 年）。これらのバイオマーカーの特異性を評価するための明確なルールが必要である。

毒物学の文脈において、最も有用であり進歩しているオミクス技術は、MOA の解析と AOP の誘導、バイオマーカーの同定であり、これらはすべて疫学にも役立つ可能性がある。例えば、(a)トランスクリプトミクス：遺伝子発現（mRNA）プロファイルの比較は、バイオマーカーの発見、発現遺伝子の機能グループ（Gene Ontology カテゴリー）へのグループ化、または遺伝子セット分析に使用することができる。これらの手法により、生物学的メカニズムに関する様々な情報が得られる可能性がある。(b)プロテオミクス：試料のタンパク質プロファイリングを調べ、タンパク質の量や転写後の修飾を高度に分析し、ばく露後の生物学的経路の変化や疾患の発症に関連している可能性がある。(c)メタボロミクスは、核磁気共鳴分光法や質量分析法をベースにした技術を用いてデータを作成し、ソフトウェアやデータベースを介して分析することで、ばく露や疾患と相関のあるマーカー（分子シグネチャーや経路）を特定するものである。(d)エクスポソーム（個人が生活の中で受けたばく露の総量）の利用は、ヒトのバイオモニタリングに適したオミクス技術とバイオマーカーを使用することで、より良い結果を得ることができるかもしれない。それにもかかわらず、これらの方法論の検証不足とそのコストの問題から、大規模な使用には限界がある。

オミクス技術を環境衛生研究に応用するには、研究デザイン、バリデーション、再現性、時間的分散、メタデータ分析に特別な配慮が必要である（Vlaanderen ら、2010 年）。大規模な研究では、生体サンプルで測定された分子プロファイルの個人内変動は、時間の経過とともに大きく変動する遺伝子発現、タンパク質レベル、または代謝物のプロファイルの個人間変動よりも少ない変動を示す。これらの個人間変動がばく露変化に関連した変動よりも大きくならないことが重要であるが、これが実現するかどうかは定かではない。

生物学的に意味のあるオミクスのシグネチャーは、オミクスーばく露及びオミクスー健康の関連性の研究を行うことによって同定され、高度なリスク評価に有用なデータを提供する。このアプローチは、末端の毒性エンドポイントから、化学物質によって誘発された分子／細胞応答の摂動に起因する毒性経路の初期の主要イベントへと移行することを支持するものである（NRC、2007 年）。

7.7. 疫学における新たなデータの機会

現在の技術的状況では、スマートフォン、テキストメッセージ、クレジットカードでの購入、オンラインでの行動、電子カ

ルテ、全地球測位システム（GPS）、スーパーマーケットの購買データなど、多くの情報源からの前例のない量のデータのデジタル化と保存が可能になっている。これらのデータ源の中には、リスク評価のための貴重な情報を提供するものもあるが、それらの多くには法的枠組みを超えて、科学的または規制上の目的のために使用することの倫理性について疑問が生じることもある。具体的な例としては、電子カルテ、職業や環境に関するアンケート、地理的な位置、健康や社会保障番号など、機密性が高い、または特に保護されていると考えられる健康に関連する個人情報を含むデータが挙げられる。これらの様々な形態の健康情報が容易に作成、保存、アクセスされている。ビッグデータは、研究者に多くのデータ源をまたいで記録を照合したり、リンクしたりする能力を提供する。健康情報と遺伝性情報のビッグデータ源のリンクは、病気の予測因子を理解するための大きな可能性を期待されている（Salerno ら、2017 年）。しかし、現在の方法を用いてデータを体系的かつ効率的に処理、解析、解釈すること、あるいは大量データの中から関連するシグナルを特定することには課題がある（全米科学・工学・医学の環境研究・毒物学委員会が 2017 年の報告書で指摘している）[18]。

さらに、国民健康保険や退院データベースから抽出された薬剤費などの医療行政データは、農業人口調査や地理的マッピングから抽出された農業活動に関するデータと相互にリンクさせることができる。このような情報が集団レベルでしか得られない場合もあることは認識されているが、個人レベル及び／または個人の習慣に関するデータを得ることが重要な課題である。

バイオバンクはまた、健康な集団や病気にかかった集団からの新たなデータ源を構成するものである。バイオバンクは、多様な研究目的のために保存されているヒトの生物学的標本と関連情報の整理されたコレクションで構成されている。これらのバイオサンプルは、ばく露評価やばく露の再構成に有用なデータを生成する可能性のある新技術の応用に利用可能である。研究の計画と実施が調和されていれば、データとサンプルはバイオバンク間で共有され、強力な蓄積分析や反復研究を促進することができる（Burton ら、2010 年）。

深い表現型（deep phenotyping）を用いた大規模な疫学研究は、優れた表現型を持つ研究参加者と前述のデータを結びつける前例のない機会を提供する。例えば、英国のバイオバンクでは、アンケート、病歴、身体測定データだけでなく、50 万人以上の参加者全員のゲノムワイドな関連データを持つ血液と尿のサンプルを保存し、病院の事例統計、国の登録データ、プライマリーケアの記録とリンクさせている。大気汚染や騒音レベルに関する情報を得るために、参加者の郵便番号を大気汚染や騒音の推定値にリンクさせている。さらに、これらのばく露に関する個人レベルのデータを収集するために、個人ばく露モニタリングの試験的実施が行われる予定である。これらのアプローチは、地理的リンク、購買・職業登録とのリンク、個人のばく露モニタリングのいずれかを通じて、農薬ばく露に関する情報を得るために拡張される可能性がある。同様のバイオバンクは、他の多くの EU 諸国にも存在する（http://www.bbmri-eric.eu/BBMRI-ERIC は、ほとんどの EU の研究を収集している）。

8. 全体的な推奨事項

8.1. 単一の疫学的研究に関する勧告

疫学的研究を改善するための以下の勧告は、規制（EU）No 1107/2009 で言及されている「認知された基準」に準拠し、農薬のリスク評価に特に価値のあるものとすることを目的としている（「利用可能で、ばく露レベルとばく露期間に関するデータを裏付けとし、認識された基準に従って実施された場合、疫学的研究は特に価値があり、提出しなければならない」としている）。したがって、これらの勧告は、このような研究をどのように実施するかについての研究者のための実践的な指針としてではなく、農薬リスク評価にさらに活用するための研究を計画している研究者のためのものと考えることができる。

a）研究デザイン（交絡を含む）

1）前向き疫学的デザインは因果関係推論のためのより強力なエビデンスを提供するので、農薬リスク評価のための他の計画よりもこれらの研究が奨励される。

[18] 全米科学・工学・医学アカデミー、地球・生命研究部門、環境研究・毒物学委員会、リスクベースの評価に 21 世紀の科学を組み込む委員会。ワシントン（DC）。全米アカデミープレス（米国）；2017 年 1 月

2）今後の疫学研究は、調査対象の課題に適切に答えるために、適切な標本数を用いて実施されるべきである。そのためには、研究デザインの段階で検出力解析を行う必要がある。

3）今後の研究では、異質性、小集団、ばく露方法、感受性の時期や条件（妊娠、発育、疾患など）を考慮する必要がある。

4）幅広い潜在的交絡変数（他の化学物質への共ばく露、ライフスタイル、社会経済的要因など）を研究の計画段階（例：マッチング）で測定または考慮すべきである。

5）毒性に影響を与え、効果を調節する宿主因子を考慮する。これらには、遺伝的多型データ（例：パラオキソナーゼ－1 遺伝子型）や栄養因子（例：ヨウ素の状態）などが含まれる。

6）研究者間の共同研究は、個々のコホートの有効性を高めるコンソーシアムを構築するために奨励される。

将来のばく露評価のために、新規技術の利用を含め、関連する生物学的試料の収集と適切な保管を行うべきである。

b）ばく露（測定、報告のためのデータ変換、統計解析）

1）ばく露に関する特定の情報の収集は、可能な限り、ばく露の広範な定義、非特定の農薬の記述及び「一度もない」対「今までにあり」のような広範なばく露の分類を避けるべきである。それにもかかわらず、これらのカテゴリーは、特定の状況下では、例えばクラス効果を予測するために価値があるかもしれない。

2）農薬の幅広いクラス（無関係な物質の一般的なグループ）、または「殺虫剤」、「除草剤」など、あるいは一般的な「農薬」だけを対象とした研究は、リスク評価にはあまり役に立たない（あるとすれば）。特定の名前のついた農薬や共調合製剤を調査した研究の方がリスク評価には有用である。

3）同じ化学クラスに属する農薬、または同じ毒性作用モードや毒性学的効果をもたらす農薬は、同じカテゴリーにグループ化されている可能性がある。ばく露の頻度、持続時間、ばく露の強度などの情報を追加することで、ばく露パターンの推定に役立てることができるかもしれない。

4）職域疫学研究では、作業者や労働者の行動や PPE の適切な使用は、これらのばく露修飾がばく露を大幅に変化させ、それによって潜在的な関連性を変化させる可能性があるため、適切に報告されなければならない。

5）ばく露測定の精度を向上させることは、特にコホート研究において、ますます重要になってきている。病因学的に関連する期間をカバーする長期のコホート研究では、繰り返しの生物学的測定や自己報告されたばく露の繰り返しの更新を使用することにより、ばく露の測定の精度を向上させるべきである。

6）農薬の使用記録、登録データ、GIS、地理的マッピングなどを含むより広い集団の環境ばく露の間接的な尺度及び大規模なデータベース（行政データベースを含む）から得られたデータは、探索的研究には貴重であるかもしれない。これらのデータが利用できない場合は、記録・登録を開始すべきである。同様に、食品消費データベースからの農薬への食事ばく露の推定や、モニタリングプログラムからの残留農薬レベルの推定も利用できる。直接的なばく露評価と同様に、間接測定の各方法は、偏りや誤分類のリスクを検討し、適切な重み付けを行うべきである。

7）可能な限り、ばく露評価は、異なるばく露レベルを確立するために、指定された農薬へのばく露の直接測定を使用すべきである（例えば、個人的なばく露測定／生物学的モニタリング）。新しい研究では、個人ばく露モニタリングの新しい方法を探求すべきである。結果は、集団間のばく露を標準化するために、標準化された単位を用いて表現されるべきである。

8）長期にわたるばく露評価の特性評価は、より包括的なばく露モニタリング戦略を実施し、アンケートやバイオモニタリングデータに裏付けられた作業ばく露マトリックスから収集された長期にわたるばく露決定要因に関する情報と相まって、利益を得ることができる。ばく露評価モデルは、重要なばく露パラメータを特定することを可能にする HBM の研究によって包括的にサポートされることができる。そのような場合には、モデル内のパラメータの仮定を調整することで、より現実的なばく露の評価につなげることができる。

9）エクスポソームの概念とメタボロミクスの使用は、特に、より良いばく露測定（ばく露のバイオマーカー）、脆弱な小集団の特定、毒性経路の生物学的解釈（疾患のバイオマーカー）のための次世代の疫学研究に大きな可能性を秘めている。

10）農薬混合物へのばく露（及び毒性）に関する知識の向上は、包括的なリスク評価に有益である。共通の標的に作

用する複数の農薬への複合ばく露の共同作用を考慮すること、または類似の有害作用を誘発することは、累積リスク評価に関連している。そのためには、混合物の全成分を把握し、MOA、用量反応特性、成分間の潜在的な相互作用を理解しておく必要がある。ばく露の特性を把握することは、ばく露のパターンや大きさが時間の経過とともに変化する複数の農薬への複合ばく露において重要な要素である。

c)有害な健康影響(測定、報告のためのデータ変換、統計解析)。

1)自己申告による健康上の影響は、研究に指定された医療専門家による病状の独立した盲検評価により、回避するか、または結論を出すべきである。

2)研究の対象となる健康影響は十分に定義されているべきであり、妥当性が確認されている場合を除き、代替エンドポイントは避けるべきである。疾患や疾患のサブクラスの定義が時間の経過とともに変化する場合には注意が必要である(がん、神経変性疾患など)。

3)早期の生物学的効果を示す生物学的マーカーを利用して、疾患の病態の理解を深めるべきである。これらの定量的な生物学的パラメータは、実験動物を用いた研究から得られた結果と比較して、研究の感度を向上させ、誤分類を減らし、ヒトへの関連性を高めることができるため、疫学の有用性を高めることができる。これらの再定義されたエンドポイントは、毒性力学的経路における初期のイベントであり、連続的なスケールで測定されることが多いため、よりあからさまな従来の健康影響よりも好ましいかもしれない。

4)効果のバイオマーカーの使用は、農薬への総体的ばく露を評価し、累積的なリスク評価に役立てることができるかもしれない。

5)健康影響を疫学研究を用いて特定し、急性・慢性の事故記録と実験結果を結びつけることを可能にするリードアクロス手法(read across methods)を開発する。

d)統計(記述的統計、ばく露と影響の関係のモデル化)。

1)統計解析は、事前に定められた解析(統計)プロトコールに基づき、探索的研究のための事後的な解析を避け、統計的に有意であるかどうかにかかわらず、すべての結果を報告すべきである。

2)データは、適切な場合には、直接または間接的な測定法が使用されているかどうかにかかわらず、個人/集団のばく露と用量反応評価を推定するための数学的モデル化を可能にするような方法で報告されなければならない。

3)報告書には、関連因子やばく露の構造が異なる基礎となる集団に基づいた研究において、未調整と調整の両方の割合と、対象となる結果の割合と率を含めるべきである。

4)考えられる関連因子及びばく露と健康転帰の関係におけるそれらの役割は、慎重に同定され、正確に測定され、徹底的に評価されるべきである。ほとんどの場合、関連因子は潜在的な交絡因子としてスクリーニングされている。交絡因子が検出された場合には、感度分析を含む適切な統計的手法を用いて、交絡因子を調整する必要がある。

5)プロペンシティスコアマッチング(propensity score matching)、メディエーション解析(mediation analyses)、因果推論などの潜在的に有用な解析手法を農薬疫学に適用することが奨励されている。

6)対象とした農薬ばく露と疾患との間の関連が統計的に有意であることが判明した場合、特に(推定される)検出力の低い研究では、統計的に有意な効果の大きさの推定値(例えば、オッズ比 OR または相対リスク RR)が人為的に増大するか、または拡大するかの程度を決定するために、検出力解析/設計計算を実行することが一般的に良い対応とされている[19]。

[19] 検出力とサンプルサイズの推奨事項及び効果量の算出とデザインの計算を含む関連問題に関する追加情報は、本報告書の附属書 D に記載されている。特に、検出力の計算では、疫学研究で明確に報告されるべき 3 つの値が必要である。(i)非ばく露群の被験者数(対象となる疾患の有無を含む)、(ii)ばく露群の被験者数(対象となる疾患の有無を含む)、(iii)非ばく露群の罹患者数。

e)結果の報告

1)結果報告は、STROBE 声明及び統計報告に関する EFSA ガイドライン(EFSA、2014 年 b)に概説されている疫学研究の良好な報告の慣行に従うべきであり、効果量の推定値を含む本意見書で指摘されている更なる提案を含むべきである。

2)いくつかの疫学研究は探索的で事後的な特性を持つものもあるがが、そのことを認識し、適切な統計解析によって裏付けられるべきである。

3)疫学研究は、さらなる調査のために生データへのアクセスを提供し、その全結果と解析に使用したスクリプトやソフトウェアパッケージを提供することが奨励されている。

4)結果の再現性を検証するために使用したスクリプトや統計ツールとともに、すべての結果を報告するか、またはオンライン源を利用して寄託する。

5)すべての資金源を報告し、財務上の問題やその他の潜在的な利害関係者を適切に報告する。

一般的な勧告として、PPR パネルは、リスク評価における疫学研究の価値、透明性、説明責任を高めるために、疫学研究のためのガイダンスの開発を奨励している[20]。 疫学研究の質の向上は、責任ある研究の実施と科学的誠実さとともに、リスク評価にこれらの研究を組み入れることに利益をもたらすであろう。

8.2. サーベイランス

1)EU 指令 2009/128 の第 7 条で要求されている市販後サーベイランスプログラム(産業用及び一般集団)を設定することにより、急性及び慢性の事故の報告を増加させる。これは、産業保健医とのサーベイランスネットワークを構築し、PPP を扱う国の当局と毒物管理情報センターとの連携を強化することにより達成すべきである。

2)急性・慢性事故の因果関係の程度/強度(「推定可能性」)を評価するための有効な方法を開発し、EU 加盟国間の整合化された報告を支援するための用語集と類義語集を開発する。

3)EU 加盟国からの整合化されたデータを EU レベルで収集し、欧州委員会/EFSA が定期的に検討し、最も関連性の高い結果に焦点を当てた報告書を発表すべきである。

4)農薬に関する EU 全体の警戒体制を構築する。

5)診断決定、データ入力、管理を担当する医療・救急スタッフのための毒物学コースにおいて、農薬のトキシドロームに関する研修を改善する余地がある。

8.3. 複数の疫学研究のメタアナリシス

1)個々の研究の方法とバイアスの徹底的な評価、研究間の異質性の程度の評価、異質性の根底にある説明の展開、エビデンスの定量的な要約(一貫性があれば)を考慮に入れて、疫学研究からのエビデンスを蓄積することができる。

2)すべてのエビデンス統合作業において、関連するバイアスのリスクツールを用いて研究をレビューすべきである。計画が異なる研究や計画の特徴が異なる研究では、バイアスのリスク評価のために(いくつかの)異なる課題が必要になるかもしれない。

3)エビデンス統合は、特定の期間に限定されるべきではなく、エビデンス全体を含めるべきである。これらの努力は、特定の健康影響または疾患カテゴリーに焦点を当てれば、より適切である。

4)エビデンス統合の取り組みでは、効果量の定量的な統合に加えて、計算された予測間隔、小試験効果と非対称性バイアス、対象となる対角線、交絡、過剰な有意性バイアス[21]及び不均一性の推定値についても考慮すべきである。

[20] 例として、オランダ疫学協会が責任ある疫学的研究実践に関するガイドラインを作成した(2017 年)

[21] 過剰シグニフィカンスバイアスとは、特定のアウトカムに関する公表されている文献の中に、統計的に有意な結果が得られた研究が多すぎる状況を指す。このパターンは、出版バイアスを伴う文献の強いバイアスを示唆している。
選択的な結果報告、選択的な分析報告、または捏造されたデータが説明の対象となる可能性がある(Ioannidis and Trikalinos、2007 年)。

5）不均質性が存在する場合、高度に選択された母集団を用いた研究は、それぞれの母集団を代表するものではないが、統計的な不均質性ではなく真の不均質性を示すものである可能性があるため、価値があることが証明され、検討に値する。

6）年齢、人種、性別など、研究間での一貫性のある報告があれば、メタアナリシスがより充実すると思われる。

7）個々の農薬の定量的データが疫学研究から得られる場合には、それらのデータを組み合わせたり、プールして用量反応モデリングを行うことで、定量的なリスク推定値や Point of Departure（BMDL、NOAEL）の開発が可能となる。

8）個々のコホートでは十分な統計的力がない疾病ばく露関連を研究するために、コホート研究の国際的なコンソーシアムがデータプールを支援するよう奨励されるべきである（例：AGRICOH）。

8.4. 疫学的証拠と他の情報源との統合

1）すべてのエビデンス（疫学、動物、試験管内データ）は、バイアスがなく、平等に精査されるべきである。

2）リスク評価のために、観察による研究、動物／基礎科学研究、その他のエビデンス源を組み合わせるために、有効かつ調和のとれた方法を開発すべきである。

3）実験データとヒトデータの両方が、ハザードの特定と用量反応評価に寄与するべきである。

4）複数のエビデンスから得られたデータを体系的に統合することは、Bradford Hill 基準を修正したものを用いて、関連性、一貫性、生物学的妥当性を考慮した WOE 解析に基づくべきである。この枠組みの基礎となる原則は 7.2 節に記載され、図 5 に要約されている。

5）WOE アプローチを用いて、疫学的知見を他の情報源（実験毒性学からのデータ、作用機序／AOP）と統合すべきである。統合された調和のとれたアプローチは、体系的かつ一貫した方法で WOE の全体的な枠組みの中で動物、メカニズム、ヒトのデータをまとめることによって開発されなければならない。

6）AOP フレームワークは、様々な種類の研究成果を統合するための構造化されたプラットフォームを提供する。

7）動物データ、in vitro データ、ヒトデータは、それぞれのエンドポイントについて全体として評価されるべきである。実験から得られた結果が各エンドポイントのヒトのデータと一致しているかどうかの結論を導き出すことができ、これを RARs に含めることができる。

9. 結論

本意見書は、既存の疫学研究を含む科学的に査読された公表文献の検索を要求する規制 1107/2009 の下で、農薬の認可更新時（可能であれば認可プロセス時）の査読プロセスを支援することを目的としている。これらは有効成分の更新プロセスに適しており、更新のために提出される書類には有効成分に関連する新しいデータを含める必要があることを示す規則 1141/2010 にも準拠している。

以下では、参照条件の 4 つの重要な要素を繰り返し、個々の用語に対応する部分を順番に示している。それぞれの ToRs でグループ化された文章から導かれるように、それぞれの ToR に関連する推奨事項も以下のように示す。

「PPR パネルは、外部科学報告書（Ntzani ら、2013 年）で観察された農薬ばく露とヒトの健康影響との関連性と、これらの知見が規制上の農薬リスク評価の文脈でどのように解釈され得るかを検討する。したがって、PPR パネルは、報告書で収集された疫学研究を体系的に評価し、研究の主要なデータギャップと限界に対処し、その提言を行う」。

PPR パネルは特に以下を行う。

1）利用可能な疫学研究の質と妥当性に関して外部科学報告書で明らかにされたものに基づいて（必ずしもこれに限定されないが）ギャップと限界のすべての情報源を収集し、レビューする。セクション 3、20-24 頁、セクション 5.2 33-35 頁：勧告は適切ではない。

2）ポイント 1 で明らかになったギャップと限界に基づき、調査結果の質、妥当性、信頼性を向上させ、それが農薬リスク評価にどのように影響を与えるかについて、将来の疫学的研究のための潜在的な改善点を提案する。これには、研究デザイン、ばく露評価、データの質とアクセス、健康影響の診断分類、統計解析が含まれる。（セクション 4 の

回答 24-33 頁:セクション 8.1、8.2 及び 8.3 の推奨事項 54-58 頁。54‐58）

3）情報及び／または基準が不十分または不足している分野を特定し、リスク評価における適用を改善し最適化するために、農薬疫学的研究をどのように実施するかについての提言を行う。これらの推奨事項には、第 1 項で特定されたギャップと限界に基づいて、ばく露評価（バイオモニタリングデータの使用を含む）、脆弱な集団のサブグループ及び／または対象となる健康影響（生化学的、機能的、形態学的、臨床的レベルでの）の調和を含むべきである。（セクション 4.2-4.5 の回答 27-33 頁、セクション 5.3 36 頁:セクション 8.1 c) 1-4、56 頁の推奨事項）

4）WOE などの評価報告書草案のピアレビュープロセスにおいて、疫学的情報を実験毒性学、AOP、作用機序などのデータと統合しながら、農薬のリスク評価に疫学的知見を適切に活用する方法について議論する。（6.2 及び 6.3 節の回答 37-45 頁、7 節 45-54 頁:8.4 節の回答 58 頁）

上記で説明したように、適切な疫学的データと承認後のサーベイランスは、ハザードの特定や、方法論の改善によりハザードの特徴を明らかにすることで、リスク評価の枠組みに有用に貢献することができる。WOE 解析、不確実性分析、バイアスの同定と推定により改善することができる。利用可能な関連文献を収集し、システマティックレビューを含む関連する EFSA 基準を用いてその妥当性と質を検討し、EU 法で規定されている DAR、RAR、承認後の枠組みの中で結果の議論を導入することは申請者の責任である。

適切な品質の定義には、サンプルサイズの解析、統計的手続き、効果の大きさの推定値の浮動、バイアスの評価、導き出された結論への寄与が必要である。研究の特性は、リスク評価プロセスのすべての関連ポイントで考慮する必要があり、例えば、生殖に関する疫学的データは、生殖影響を明らかにするために計画された実験動物研究と一緒に考慮され、生殖毒性（ECHA の場合）のための表示勧告の背景で考慮される。

EU 域外の国での使用実績がない限り、関連する疫学的研究は、DAR への影響が制限されるが、RAR とサーベイランスの枠組みは、最初の承認後の時間が経つにつれて、また他の法域での原薬の先行使用がある場合には、段階的に疫学から恩恵を受けることができる可能性がある。RAR とサーベイランスのプロトコールはこの違いを反映させることが推奨される。

上記の提言は、リスク評価における疫学的データの利用に関連した現在及び将来の強み・弱み・機会・脅威の解析に基づく詳細な論拠に基づくものである。大まかには以下の通りである。

強み:

- その証拠は人間の特定のリスクに関するものであること。
- 健康影響は、毒素へのすべてのばく露の影響を総合的に測定するものであること。
- 影響を受ける可能性のある人々から主観的な経験を引き出すことができること。

弱み:

- 農薬へのばく露は通常複雑である;特定の有効成分の寄与は容易に解読されない。
- ばく露は、正確に管理された条件が不足している様々な設定で発生する。
- ほとんどのデータは混合集団の反応を再現している。
- 多くのデータは低レベルの関連性を示しており、再現性がなく、高度な解析を必要とする。

機会:本意見書に記載されている限界は、公表されている多くの疫学研究に適用されるが、農薬のリスク評価に利益をもたらす機会がある。これには以下が含まれる。

- 軽微な健康影響を明らかにし、敏感な小集団の経験を明らかにする可能性のある研究のために、非常に多くの潜在的にばく露された個人にアクセスできること。
- 潜在的な毒素とその残留物の組織負荷を確立するためのバイオモニタリングと新しい分子アプローチを用いたばく露推定の改善の見通し。
- 実験動物の反応に基づく従来のリスク評価にヒトのデータを完全に統合する可能性。
- WOE, AOP, Expert judgement, Expert Knowledge Elicitation（EEKE）及び Uncertainty Analysis を利用して、潜在的に関連性のあるデータの質の違いを評価する。
- 専門の疫学者や統計学者と協力して、疫学的結果の解釈を再検討し、慢性ばく露リスクや複合ばく露リスク、用量

反応データなどの困難な分野に取り組むための改善案を提案する機会がある。

- 様々な国の情報源からのデータを蓄積することには、大きな情報技術の機会が存在する。関連する法的、方法論的、倫理的な問題が克服されれば、より多くの価値あるデータを収集することができる。このデータが社会的利益のために「ビッグデータ」の設定で使用できる形で利用できるようになれば、疫学研究を大幅に改善できる可能性がある。しかし、第一に、個人のプライバシーと本質的な商業的な機密性を守る必要がある。これらの障害が克服されれば、疫学研究の統計的な力を向上させ、ハザードをより良く特定し、場合によっては特徴づけるために応用することができる。これらの目的は、EU の高いレベルで合意された行動によって効果的に実現することができる。データとインタラクティブなプラットフォームを提供するための相互間承認は、集団の健康情報、食品消費データ、有効成分と共調合製剤の空間的・時間的な適用データの調和によって裏付けられる必要がある。このような豊富なデータは、因果関係と信頼性のエビデンスを強化する基準である一貫性の向上を支援することが期待できる。それは、農薬毒性からの特別な保護を必要とするかもしれない脆弱なグループをよりよく特定できるようになる疫学研究のためのより大きなサンプルサイズを約束している。

脅威:

- 非現実的であり、社会に否定的な結果をもたらすヒト集団や野生生物、環境に対するリスクレベルの認識が広まっていること。
- 他の有効な情報源からのデータを損なうような、誤った陽性または誤った陰性の結論をもたらす不十分な実験計画。
- 効果的なサーベイランスが行われていなかったり、匿名化された適切なデータを社会的利益のために利用できるようにする気がなかったりした結果、新たなリスクに対応できなかったこと。
- 登録（がんや先天性異常）やサーベイランスプログラムによるばく露（特に職業上ばく露）に関する適切な情報収集の失敗によるデータの浪費。
- 診断基準の調和の失敗、統合解析のための十分に詳細な組み合わせ可能な形でデータを記録しなかったことによるデータの浪費、健康統計データベースに入力されるデータの最適な品質を可能にするための関連するトキシドロームについての医療や救急医療スタッフのトレーニング不足。

参考文献

Adami HO, Berry SC, Breckenridge CB, Smith LL, Swenberg JA, Trichopoulos D, Weiss NS and Pastoor TP, 2011. Toxicology and epidemiology: improving the science with a framework for combining toxicological and epidemiological evidence to establish causal inference. Toxicology Sciences, 122, 223–234.

Amler RW, Barone Jr S, Belger A, Berlin Jr CM, Cox C, Frank H, Goodman M, Harry J, Hooper SR, Ladda R, LaKind JS, Lipkin PH, Lipsitt LP, Lorber MN, Myers G, Mason AM, Needham LL, Sonawane B, Wachs TD and Yager JW, 2006. Hershey Medical Center Technical Workshop Report: optimizing the design and interpretation of epidemiologic studies for assessing neurodevelopmental effects from in utero chemical exposure. Neurotoxicology, 27, 861–874.

Bengtson AM, Westreich D, Musonda P, Pettifor A, Chibwesha C, Chi BH, Vwalika B, Pence BW, Stringer JS and Miller WC, 2016. Multiple overimputation to address missing data and measurement error: application to HIV treatment during pregnancy and pregnancy outcomes. Epidemiology, 27, 642–650.

Bevan R, Brown T, Matthies F, Sams C, Jones K, Hanlon J and La Vedrine M, 2017. Human Biomonitoring data collection from occupational exposure to pesticides. EFSA supporting publication 2017: EN-1185, 207 pp.

Bottai M, 2014. Lessons in biostatistics: inferences and conjectures about average and conditional treatment effects in randomized trials and observational studies. Journal of Internal Medicine, 276, 229–237.

Budtz-Jørgensen E, Keiding N and Grandjean P, 2001. Benchmark dose calculation from epidemiological data. Biometrics, 57, 698–706.

Budtz-Jørgensen E, Keiding N and Grandjean P, 2004. Effects of exposure imprecision on estimation of the benchmark dose. Risk Analysis, 24, 1689–1696.

Buonsante VA, Muilerman H, Santos T, Robinson C and Tweedale AC, 2014. Risk assessment's insensitive toxicity testing may cause it to fail. Environmental Research, 135, 139–147.

Burton PR, Fortier I and Knoppers BM, 2010. The global emergence of epidemiological biobanks: opportunities and challenges. In: Khoury M, Bedrosian S, Gwinn M, Higgins J, Ioannidis J and Little J (eds.). Human Genome Epidemiology. Building the evidence for using genetic information to improve

health and prevent disease. 2nd Edition, Oxford University Press, Oxford. pp. 77–99.
Choi J, Polcher A and Joas A, 2016. Systematic literature review on Parkinson′s disease and Childhood Leukaemia and mode of actions for pesticides. EFSA supporting publication 2016:EN-955, 256 pp. Available online: http:// www.onlinelibrary.wiley.com/doi/10.2903/sp.efsa.2016.EN-955/pdf
Coble J, Thomas KW, Hines CJ, Hoppin JA, Dosemeci M, Curwin B, Lubin JH, Beane Freeman LE, Blair A, Sandler DP and Alavanja MC, 2011. An updated algorithm for estimation of pesticide exposure intensity in the agricultural health study. International Journal of Environmental Research and Public Health, 8, 4608–4622.
Coggon D, 1995. Questionnaire based exposure assessment methods. Science of the Total Environment, 168, 175–178.
Cornelis C, Schoeters G, Kellen E, Buntinx F and Zeegers M, 2009. Development of a GIS-based indicator for environmental pesticide exposure and its application to a Belgian case-control study on bladder cancer. International Journal of Hygiene and Environmental Health, 212, 172–185.
la Cour JL, Brok J and Gøtzsche PC, 2010. Inconsistent reporting of surrogate outcomes in randomised clinical trials: cohort study. BMJ, 341, c3653.
DeBord DG, Burgoon L, Edwards SW, Haber LT, Kanitz MH, Kuempel E, Thomas RS and Yucesoy B, 2015. Systems biology and biomarkers of early effects for occupational exposure limit setting. The Journal of Occupational and Environmental Hygiene, 12(Suppl 1), S41–S54.
Dionisio KL, Chang HH and Baxter LK, 2016. A simulation study to quantify the impacts of exposure measurement error on air pollution health risk estimates in copollutant time-series models. Environmental Health, 15, 114.
DSE (Dutch Society for Epidemiology), 2017. Responsible Epidemiologic Research Practice (RERP). A guideline developed by the RERP working group of the Dutch Society for Epidemiology, 2017 (available at https:// www.epidemiologie.nl/home.html, https://epidemiologie.nl/fileadmin/Media/docs/ Onderzoek/ Responsible_Epide miologic_Research_Practice.2017.pdf)
ECETOC (European Centre for Ecotoxicology and Toxicology of Chemicals), 2009. Framework for the Integration of Human and Animal Data in Chemical Risk Assessment. Technical Report No. 104. Brussels. Available online: http://www.ecetoc.org/uploads/Publications/documents/TR%20104.pdf
ECHA/EFSA, 2014. Workshop on Mode of action and Human relevance framework in the context of classification and labelling (CLH) and regulatory assessment of biocides and pesticides. November 2014. Available online: https://echa.europa.eu/documents/10162/22816050/moaws_ workshop_proceedings_ en.pdf/ a656803e-4d97-438f-87ff-fc984cfe4836
EFSA (European Food Safety Authority), 2004. Opinion of the Scientific Panel on Dietetic Products, Nutrition and Allergies on a request from the Commission related to the presence of trans fatty acids in foods and the effect on human health of the consumption of trans fatty acids. EFSA Journal 2004;81, 1–49 pp. https://doi. org/10.2903/j.efsa.2004.81
EFSA (European Food Safety Authority), 2009a. Scientific Opinion of the Panel on Contaminants in the Food Chain on a request from the European Commission on cadmium in food. EFSA Journal 2009;980, 1–139 pp. https:// doi.org/10.2903/j.efsa.2009.980
EFSA (European Food Safety Authority Panel on Contaminants in the Food Chain CONTAM), 2009b. Scientific Opinion on arsenic in food. EFSA Journal 2009;7(10):1351, 199 pp. https://doi.org/10.2903/j.efsa.2009.1351
EFSA (European Food Safety Authority), 2010a. Application of systematic review methodology to food and feed safety assessments to support decision making. EFSA Journal 2010;8(6):1637, 90 pp. https://doi.org/10.2903/ j.efsa.2010.1637
EFSA (European Food Safety Authority) Panel on Contaminants in the Food Chain (CONTAM), 2010b. Scientific Opinion on Lead in Food. EFSA Journal 2010;8(4):1570, 151 pp. https://doi.org/10.2903/j.efsa.2010.1570
EFSA (European Food Safety Authority), 2011a. Submission of scientific-peer reviewed open literature for the approval of pesticide active substances under Regulation (EC) No 1107/2009. EFSA Journal 2011;9(2):2092, 49 pp. https://doi.org/10.2903/j.efsa.2011.2092
EFSA (European Food Safety Authority), 2011b. Statistical significance and biological relevance. EFSA Journal 2011;9(9):2372, 17 pp. https://doi.org/10.2903/j.efsa.2011.2372
EFSA (European Food Safety Authority), 2012a. Scientific Opinion on risk assessment terminology. EFSA Journal 2012;10(5):2664, 43 pp. https://www.efsa.europa.eu/en/efsajournal/pub/2664
EFSA (European Food Safety Authority Panel on Contaminants in the Food Chain CONTAM), 2012b. Scientific Opinion on the risk for public health related to the presence of mercury and methylmercury in food. EFSA Journal 2012;10(12):2985, 241 pp. https://doi.org/10.2903/j.efsa.2012.2985
EFSA (European Food Safety Authority), 2013a. Scientific Opinion on the identification of pesticides to be included in cumulative assessment groups on the basis of their toxicological profile. EFSA Journal 2013;11(7):3293, 131 pp. https://doi.org/10.2903/j.efsa.2013.3293
EFSA (European Food Safety Authority), 2013b. Scientific Opinion on the relevance of dissimilar mode of action and its appropriate application for cumulative risk assessment of pesticides residues in food. EFSA Journal 2013;11(12):3472, 40 pp. https://doi.org/10.2903/j.efsa.2013.3472
EFSA (European Food Safety Authority), 2014a. Conclusion on the peer review of the pesticide human health risk assessment of the active substance chlorpyrifos. EFSA Journal 2014;12(4):3640, 34 pp. https://doi.org/ 10.2903/j.efsa.2014.3640
EFSA (European Food Safety Authority), 2014b. Guidance on statistical reporting. EFSA Journal 2014;12(12): 3908, 18 pp. https://doi.org/10.2903/j.efsa.2014.3908
EFSA (European Food Safety Authority), 2015a. Stakeholder Workshop on the use of epidemiological data

in pesticide risk assessment. EFSA supporting publication 2015:EN-798, 8 pp. Available online: https://www.efsa. europa.eu/en/supporting/pub/798e

EFSA (European Food Safety Authority), 2015b. Increasing robustness, transparency and openness of scientific assessments – Report of the Workshop held on 29–30 June 2015 in Brussels. EFSA supporting publication 2015: EN-913. 29 pp. Available online: http://www.efsa.europa.eu/sites/default/files/corporate _publications/ files/913e.pdf

EFSA (European Food Safety Authority), 2015c. Conclusion on the peer review of the pesticide risk assessment of the active substance glyphosate. EFSA Journal 2015;13(11):4302, 107 pp. https://doi.org/10.2903/j.efsa.2015.4302

EFSA PPR Panel (European Food Safety Authority Panel on Plant Protection Products and their Residues), 2017. Scientific Opinion on the investigation into experimental toxicological properties of plant protection products having a potential link to Parkinson's disease and childhood leukaemia. EFSA Journal 2017;15(3):4691, 325 pp. https://doi.org/10.2903/j.efsa.2017.4691

EFSA Scientific Committee (European Food Safety Authority Scientific Committee), 2017a. Guidance on the assessment of the biological relevance of data in scientific assessments. EFSA Journal 2017;15(8):4970, 73 pp. https://doi.org/10.2903/j.efsa.2017.4970

EFSA Scientific Committee (European Food Safety Authority Scientific Committee), 2017b. Guidance on the use of the weight of evidence approach in scientific assessments. EFSA Journal 2017;15(8):4971, 69 pp. https://doi. org/10.2903/j.efsa.2017.4971.

EFSA Scientific Committee (European Food Safety Authority Scientific Committee), 2017c. Update: guidance on the use of the benchmark dose approach in risk assessment. EFSA Journal 2017;15(1): 4658, 41 pp. https://doi. org/10.2903/j.efsa.2017.4658

von Elm E, Altman DG, Egger M, Pocock SJ and Gøtzsche PC, Vandenbroucke JP and STROBE Initiative, 2007. Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. BMJ, 335, 806–808.

Esch EW, Bahinski A and Huh D, 2015. Organs-on-chips at the frontiers of drug discovery. Nature Reviews. Drug Discovery, 14, 248–260.

Fedak KM, Bernal A, Capshaw ZA and Gross S, 2015. Applying the Bradford Hill criteria in the 21st century: how data integration has changed causal inference in molecular epidemiology. Emerging Themes in Epidemiology, 30, 14.

Gibson SB, Downie JM, Tsetsou S, Feusier JE, Figueroa KP, Bromberg MB, Jorde LB and Pulst SM, 2017. The evolving genetic risk for sporadic ALS. Neurology, 89, 226–233.

Gómez-Martín A, Hernández AF, Martínez-González LJ, González-Alzaga B, Rodríguez-Barranco M, Lopez-Flores I, Aguilar-Garduno C and Lacasana M, 2015. Polymorphisms of pesticide-metabolizing genes in children living in intensive farming communities. Chemosphere, 139, 534–540.

González-Alzaga B, Hernández AF, Rodríguez-Barranco M, Gómez I, Aguilar-Garduño C, López-Flores I, Parrón T and Lacasaña M, 2015. Pre- and postnatal exposures to pesticides and neurodevelopmental effects in children living in agricultural communities from South-Eastern Spain. Environment International, 85, 229–237.

Greenland S and Longnecker MP, 1992. Methods for trend estimation from summarized dose-response data, with applications to meta-analysis. American Journal of Epidemiology, 135, 1301–1309.

Greenland S and O'Rourke K, 2008. Meta-analysis. In: Rothman K, Greenland S and Lash T (eds). *Modern Epidemiology*. *3*. Lippincott Williams & and Wilkins, Philadelphia. pp. 652–682.

Greenland S, Senn SJ, Rothman KJ, Carlin JB, Poole C, Goodman SN and Altman DG, 2016. Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. European Journal of Epidemiology, 31, 337–350.

Grimes DA and Schulz KF, 2005. Surrogate end points in clinical research: hazardous to your health. Obstetrics and Gynecology, 105, 1114–1118.

Gustafson P and McCandless LC, 2010. Probabilistic approaches to better quantifying the results of epidemiologic studies. International Journal of Environmental Research and Public Health, 7, 1520–1539.

Hernández AF, González-Alzaga B, López-Flores I and Lacasaña M, 2016. Systematic reviews on neurodevelopmental and neurodegenerative disorders linked to pesticide exposure: methodological features and impact on risk assessment. Environment International, 92–93, 657–679.

Higgins JP, 2008. Commentary: heterogeneity in meta-analysis should be expected and appropriately quantified. International Journal of Epidemiology, 37, 1158–1160.

Hill AB, 1965. The environment and disease: association or causation? Proceedings of the Royal Society of Medicine, 58, 295–300.

Hines CJ, Deddens JA, Coble J, Kamel F and Alavanja MC, 2011. Determinants of captan air and dermal exposures among orchard pesticide applicators in the Agricultural Health Study. Annals of Occupational Hygiene, 55, 620–633.

Hoffmann S, de Vries RBM, Stephens ML, Beck NB, Dirven HAAM, Fowle JR 3rd, Goodman JE, Hartung T, Kimber I, Lalu MM, Thayer K, Whaley P, Wikoff D and Tsaioun K, 2017. A primer on systematic reviews in toxicology. Archives of Toxicology, 91, 2551–2575.

Höfler M, 2005. The Bradford Hill considerations on causality: a counterfactual perspective. Emerging Themes in Epidemiology, 2, 11.

IEA (International Epidemiological Association), 2007. Good Epidemiological Practice (GEP) 2007. Available online: http://ieaweb.org/good-epidemiological-practice-gep/

Imbens G and Rubin D, 2015. *Causal Inference for Statistics, Social, and Biomedical Sciences: An*

Introduction. Cambridge University Press, New York, NY.
INSERM, 2013. Pesticides. Effets sur la santé. Collection expertise collective, Inserm, Paris, 2013.
Ioannidis JP and Trikalinos TA, 2007. An exploratory test for an excess of significant findings. Clinical Trials, 4, 245–253.
Jurek AM, Greenland S, Maldonado G and Church TR, 2005. Proper interpretation of non-differential misclassification effects: expectations vs observations. International Journal of Epidemiology, 34, 680–687.
Kaltenhäuser J, Kneuer C, Marx-Stoelting P, Niemann L, Schubert J, Stein B and Solecki R, 2017. Relevance and reliability of experimental data in human health risk assessment of pesticides. Regulatory Toxicology and Pharmacology, 88, 227–237.
Karabatsos G, Talbott E and Walker SG, 2015. A Bayesian nonparametric meta-analysis model. Research Synthesis Methods, 6, 28–44.
Kavvoura FK, Liberopoulos G and Ioannidis JP, 2007. Selection in reported epidemiological risks: an empirical assessment. PLoS Medicine, 4, e79.
Lachenmeier DW, Kanteres F and Rehm J, 2011. Epidemiology-based risk assessment using the benchmark dose/margin of exposure approach: the example of ethanol and liver cirrhosis. International Journal of Epidemiology, 40, 210–218.
LaKind JS, Sobus JR, Goodman M, Barr DB, Furst P, Albertini RJ, Arbuckle TE, Schoeters G, Tan YM, Teequarden J, Tornero-Velez R and Weisel CP, 2014. A proposal for assessing study quality: biomonitoring, environmental epidemiology, and short-lived chemicals (BEES-C) instrument. Environmental International, 73, 195–207.
LaKind JS, Goodman M, Barr DB, Weisel CP and Schoeters G, 2015. Lessons learned from the application of BEES-C: systematic assessment of study quality of epidemiologic research on BPA, neurodevelopment, and respiratory health. Environment International, 80, 41–71.
Landgren O, Kyle RA, Hoppin JA, Beane Freeman LE, Cerhan JR, Katzmann JA, Rajkumar SV and Alavanja MC, 2009. Pesticide exposure and risk of monoclonal gammopathy of undetermined significance in the Agricultural Health Study. Blood, 113, 6386–6391.
Larsson MO, Nielsen VS, Brandt CØ, Bjerre N, Laporte F and Cedergreen N, 2017. Quantifying dietary exposure to pesticide residues using spraying journal data. Food and Chemical Toxicology, 105, 407–428.
Lash TL, Fox MP and Fink AK, 2009. *Applying Quantitative Bias Analysis to Epidemiologic Data*. Springer, New York.
Lavelle KS, Robert Schnatter A, Travis KZ, Swaen GM, Pallapies D, Money C, Priem P and Vrijhof H, 2012. Framework for integrating human and animal data in chemical risk assessment. Regulatory Toxicology and Pharmacology, 2012; 62, 302–312.
London L, Coggon D, Moretto A, Westerholm P, Wilks MF and Colosio C, 2010. The ethics of human volunteer studies involving experimental exposure to pesticides: unanswered dilemmas. Environmental Health, 18, 50.
Maldonado G and Greenland S, 2002. Estimating causal effects. International Journal of Epidemiology, 31, 422–429.
Marx-Stoelting P, Braeuning A, Buhrke T, Lampen A, Niemann L, Oelgeschlaeger M, Rieke S, Schmidt F, Heise T, Pfeil R and Solecki R, 2015. Application of omics data in regulatory toxicology: report of an international BfR expert workshop. Archives of Toxicology, 89, 2177–2184.
McNamee R, 2003. Confounding and confounders. Occupational and Environmental Medicine, 60, 227–234.
Monson R, 1990. *Occupational Epidemiology*, 2nd Edition. CRC Press, Boca Ration, FL.
Muñoz-Quezada MT, Lucero BA, Barr DB, Steenland K, Levy K, Ryan PB, Iglesias V, Alvarado S, Concha C, Rojas E and Vega C, 2013. Neurodevelopmental effects in children associated with exposure to organophosphate pesticides: a systematic review. Neurotoxicology, 39, 158–168.
Nachman KE, Fox MA, Sheehan MC, Burke TA, Rodricks JV and Woodruff TJ, 2011. Leveraging epidemiology to improve risk assessment. Open Epidemiology Journal, 4, 3–29.
Nieuwenhuijsen MJ, 2015. Exposure assessment in environmental epidemiology. In: Vrijheid M (ed.). The Exposome-Concept and Implementation in Birth Cohorts Chapter 14. Oxford University Press.
NRC (National Research Council), 2007. Toxicity Testing in the 21st Century: A Vision and a Strategy. Washington, DC: The National Academies Press.
NRC (National Research Council), 2009. *Science and Decisions: Advancing Risk Assessment*. The National Academies Press, Washington, DC.
Ntzani EE, Chondrogiorgi M, Ntritsos G, Evangelou E and Tzoulaki I, 2013. Literature review on epidemiological studies linking exposure to pesticides and health effects. EFSA supporting publication 2013: EN-497, 159 pp.
OECD (Organisation for Economic Co-operation and Development), 2013. Guidance Document on Developing and Assessing Adverse Outcome Pathways. Series on Testing and Assessment, No. 184. Paris. Avilable online: http:// search.oecd.org/officialdocuments/displaydocumentpdf/?cote=env/jm/mono%282013%296&doclanguage=en
Orford R, Crabbe H, Hague C, Schaper A and Duarte-Davidson R, 2014. EU alerting and reporting systems for potential chemical public health threats and hazards. Environment International, 72, 15–25.
Orford R, Hague C, Duarte-Davidson R, Settimi L, Davanzo F, Desel H, Pelclova D, Dragelyte G, Mathieu-Nolf M, Jackson G and Adams R, 2015. Detecting, alerting and monitoring emerging chemical health threats: ASHTIII. European Journal of Public Health, 25(supp 3), 218.
Orsini N, Li R, Wolk A, Khudyakov P and Spiegelman D, 2012. Meta-analysis for linear and nonlinear dose-response relations: examples, an evaluation of approximations, and software. American Journal of

Epidemiology, 175, 66–73.
Oulhote Y and Bouchard MF, 2013. Urinary metabolites of organophosphate and pyrethroid pesticides and behavioral problems in Canadian children. Environmental Health Perspectives, 121, 1378–1384.
Pearce N, 2011. Registration of protocols for observational research is unnecessary and would do more harm than good. Occupational and Environmental Medicine, 68, 86–88.
Pearce N, 2012. Classification of epidemiological study designs. International Journal of Epidemiology, 41, 393–397. Pearce N, Blair A, Vineis P, Ahrens W, Andersen A, Anto JM, Armstrong BK, Baccarelli AA, Beland FA, Berrington A, Bertazzi PA, Birnbaum LS, Brownson RC, Bucher JR, Cantor KP, Cardis E, Cherrie JW, Christiani DC, Cocco P, Coggon D, Comba P, Demers PA, Dement JM, Douwes J, Eisen EA, Engel LS, Fenske RA, Fleming LE, Fletcher T, Fontham E, Forastiere F, Frentzel-Beyme R, Fritschi L, Gerin M, Goldberg M, Grandjean P, Grimsrud TK, Gustavsson P, Haines A, Hartge P, Hansen J, Hauptmann M, Heederik D, Hemminki K, Hemon D, Hertz-Picciotto I, Hoppin JA, Huff J, Jarvholm B, Kang D, Karagas MR, Kjaerheim K, Kjuus H, Kogevinas M, Kriebel D, Kristensen P, Kromhout H, Laden F, Lebailly P, LeMasters G, Lubin JH, Lynch CF, Lynge E, 't Mannetje A, McMichael AJ, McLaughlin JR, Marrett L, Martuzzi M, Merchant JA, Merler E, Merletti F, Miller A, Mirer FE, Monson R, Nordby KC, Olshan AF, Parent ME, Perera FP, Perry MJ, Pesatori AC, Pirastu R, Porta M, Pukkala E, Rice C, Richardson DB, Ritter L, Ritz B, Ronckers CM, Rushton L, Rusiecki JA, Rusyn I, Samet JM, Sandler DP, de Sanjose S, Schernhammer E, Costantini AS, Seixas N, Shy C, Siemiatycki J, 2015. Silverman DT, Simonato L, Smith AH, Smith MT, Spinelli JJ, Spitz MR, Stallones L, Stayner LT, Steenland K, Stenzel M, Stewart BW, Stewart PA, Symanski E, Terracini B, Tolbert PE, Vainio H, Vena J, Vermeulen R, Victora CG, Ward EM, Weinberg CR, Weisenburger D, Wesseling C, Weiderpass E, Zahm SH. IARC monographs: 40 years of evaluating carcinogenic hazards to humans. Environmental Health Perspectives, 123, 507–514.
Raffaele KC, Vulimiri SV and Bateson TF, 2011. Benefits and barriers to using epidemiology data in environmental risk. The Journal of Epidemiology, 4, 99–105.
Raphael K, 1987. Recall bias: a proposal for assessment and control. International Journal of Epidemiology, 16, 167–170.
Rappaport SM, 2012. Biomarkers intersect with the exposome. Biomarkers, 17, 483–489.
Reich CG, Ryan PB and Schuemie MJ, 2013. Alternative outcome definitions and their effect on the performance of methods for observational outcome studies. Drug Safety, 36(Suppl 1), S181–S193.
Rothman KJ, 2002. *Epidemiology – An Introduction*. Oxford University Press, Oxford.
Rothman KJ and Greenland S, 1998. *Modern Epidemiology. 2.* Philadelphia: Lippincott Williams & Wilkins, 27 pp.
Rothman KJ, Greenland S and Lash TL, 2008. *Modern Epidemiology, 3rd Edition*. Lippincott Williams & Wilkins, Philadelphia, PA, USA.
Rushton L, 2011. Should protocols for observational research be registered? Occupational and Environmental Medicine, 68, 84–86.
Salerno J, Knoppers BM, Lee LM, Hlaing WW and Goodman KW, 2017. Ethics, big data and computing in epidemiology and public health. Annals of Epidemiology, 27, 297–301. https://doi.org/10.1016/ j. annepidem. 2017.05.002
Santacatterina M and Bottai M, 2015. Inferences and conjectures in clinical trials: a systematic review of generalizability of study findings. Journal of Internal Medicine, 279, 123–126. https://doi.org/10.1111/ joim.12389
SCENIHR, 2012. Memorandum on the use of the scientific literature for human health risk assessment purposes –weighing of evidence and expression of uncertainty.
Simera I, Moher D, Hoey J, Schulz KF and Altman DG, 2010. A catalogue of reporting guidelines for health research. European Journal of Clinical Investigation, 40, 35–53.
Skelly AC, 2011. Probability, proof, and clinical significance. Evidence-Based Spine-Care Journal, 2, 9–11.
Spiegelman D, 2016. Evaluating Public Health Interventions: 4. the nurses' health study and methods for eliminating bias attributable to measurement error and misclassification. American Journal of Public Health, 106, 1563–1566.
Stang PE, Ryan PB, Dusetzina SB, Hartzema AG, Reich C, Overhage JM and Racoosin JA, 2012. Health outcomes of interest in observational data: issues in identifying definitions in the literature. Health Outcomes Research in Medicine, 3, e37–e44.
Thomas DC, 2009. *Statistical Methods in Environmental Epidemiology.* Oxford University Press, Oxford, UK.
Thomas KW, Dosemeci M, Coble JB, Hoppin JA, Sheldon LS, Chapa G, Croghan CW, Jones PA, Knott CE, Lynch CF, Sandler DP, Blair AE and Alavanja MC, 2010. Assessment of a pesticide exposure intensity algorithm in the agricultural health study. Journal of Exposure Science & Environmental Epidemiology, 20, 559–569.
Tsilidis KK, Panagiotou OA, Sena ES, Aretouli E, Evangelou E, Howells DW, Al-Shahi Salman R, Macleod MR and Ioannidis JP, 2013. Evaluation of excess significance bias in animal studies of neurological diseases. PLoS Biology, 11, e1001609.
Turner MC, Wigle DT and Krewski D, 2010. Residential pesticides and childhood leukemia: a systematic review and meta-analysis.
US EPA (United States Environmental Protection Agency), 2011. Chlorpyrifos: preliminary human health risk assessment for registration review, 30 June 2011, 159 pp.
US-EPA (U.S. Environmental Protection Agency), 2010a. Framework for incorporating human epidemiologic & incident data in health risk assessment (draft). Office of Pesticide Programs. Washington, DC, 2010.

US-EPA (U.S. Environmental Protection Agency), 2010b. Meeting Minutes of the FIFRA Scientific Advisory Panel Meeting on the Draft Framework and Case Studies on Atrazine, Human Incidents, and the Agricultural Health Study: Incorporation of Epidemiology and Human Incident Data into Human Health Risk Assessment. Arlington, Virginia, USA, April 22, 2010b. Available online: https://archive.epa.gov/scipoly/sap/meetings/web/ pdf/020210minutes.pdf
US-EPA (U.S. Environmental Protection Agency), 2012. Guidance for considering and using open literature toxicity studies to support human health risk assessment. Office of Pesticide Programs. Washington, DC, 2012. Available online: http://www.epa.gov/pesticides/science/lit-studies.pdf
US-EPA (Environmental Protection Agency), 2016. Office of Pesticide Programs' Framework for Incorporating Human Epidemiologic & Incident Data in Risk Assessments for Pesticides December 28, 2016. Avilable online: https://www3.epa.gov/pesticides/EPA-HQ-OPP-2008-0316-DRAFT-0075.pdf
Vandenberg LN, Ågerstrand M, Beronius A, Beausoleil C, Bergman Å, Bero LA, Bornehag CG, Boyer CS, Cooper GS, Cotgreave I, Gee D, Grandjean P, Guyton KZ, Hass U, Heindel JJ, Jobling S, Kidd KA, Kortenkamp A, Macleod MR, Martin OV, Norinder U, Scheringer M, Thayer KA, Toppari J, Whaley P, Woodruff TJ and Rude, n C, 2016. A proposed framework for the systematic review and integrated assessment (SYRINA) of endocrine disrupting chemicals. Environmental Health, 15, 74.
van den Brandt P, Voorrips L, Hertz-Picciotto I, Shuker D, Boeing H, Speijers G, Guittard C, Kleiner J, Knowles M, Wolk A and Goldbohm A, 2002. The contribution of epidemiology. Food and Chemical Toxicology, 40, 387–424.
Vinken M, 2013. The adverse outcome pathway concept: a pragmatic tool in toxicology. Toxicology, 312, 158–165.
Vlaanderen J, Moore LE, Smith MT, Lan Q, Zhang L, Skibola CF, Rothman N and Vermeulen R, 2010. Application of OMICS technologies in occupational and environmental health research: current status and projections. Occupational and Environmental Medicine, 67, 136–43.
WHO/IPCS (World Health Organization/International Programme on Chemical Safety), 2009. EHC 240: principles and methods for the risk assessment of chemicals in food.
Wilson SJ and Tanner-Smith EE, 2014. Meta-analysis in prevention science. In: Sloboda Z and Petras H (eds.). *Defining prevention science*. Advances in Prevention Science (vol. 1): Defining Prevention Science Springer, New York. pp. 431–452.
Youngstrom E, Kenworthy L, Lipkin PH, Goodman M, Squibb K, Mattison DR, Anthony LG, Makris SL, Bale AS, Raffaele KC and LaKind JS, 2011. A proposal to facilitate weight-of-evidence assessments: harmonization of Neurodevelopmental Environmental Epidemiology Studies (HONEES). Neurotoxicology and Teratology, 33, 354–359.
Zingone A and Kuehl WM, 2011. Pathogenesis of monoclonal gammopathy of undetermined significance and progression to multiple myeloma. Seminars in Hematology, 48, 4–12.

用語集と略語

ADI	一日の許容摂取量。食品または飲料水に含まれる農薬の量の尺度で、相当な健康リスクを伴わずに生涯にわたって日常的に(経口的に)摂取することができる。
ADME	薬理学(及び毒物学)で使用される略語(体内動態)で、化学物質の吸収、分布、代謝及び排泄のために使用され、生物体内でのその処理を示す。
AOP Adverse Outcome Pathway	(有害性転帰経路)。リスク評価に関連する有害な影響につながる生物学的事象を構造的に表現したもの。
ARfD	急性参照用量(Acute Reference Dose)。食品または飲料水に含まれる農薬の量(通常は体重ベースで表される)の推定値で、評価時に知られているすべての事実に基づいて、24 時間以内に消費者が健康上のリスクを認めずに摂取できる量。
バイオマーカー	「生物学的マーカー」とも呼ばれる。正常な生物学的プロセス、病原性プロセス、または治療的介入に対する薬理学的反応の指標として客観的に測定され、評価される特性。
BMD	ベンチマークドーズ。バックグラウンドと比較して有害な影響の反応率(ベンチマーク反応またはBMR)に所定の変化をもたらす閾値の用量または濃度。95%の下限値(BMDL)が計算され、健康に基づいた参照値を導き出すための出発点としてさらに使用される。
HBM	ヒューマンバイオモニタリング(Human biomonitoring)。ヒトの生物学的体液または組織における化学物質及び/またはその代謝物の測定。また、すべてのばく露経路からの総合的なばく露から得られる化学物質の内部ばく露量とも呼ばれる。
ヒトデータ	研究者が研究参加者に働きかけることなく、要因と健康影響との間の自然な関係を観察する観察

による研究（疫学研究とも呼ばれる）が含まれる。警戒データもこの概念に該当する。対照的に、研究者が研究デザインの一部として介入する介入研究（実験研究または無作為化臨床試験とも呼ばれる）は、本意見書の範囲外である。

IARC International Agency for Research on Cancer（国際がん研究機関）。世界のがんの原因と発生に関する研究を実施し、調整することを役割とする世界保健機関（WHO）の機関。

LOAEL（LOAEL） 最小中毒量（Lowestobserved-adverse-effect level）。毒性試験で評価され、有害な影響（対象生物の形態、生化学、機能、または生涯への有害な変化など）を示す化学的ストレス因子の最低濃度または用量。

NOAEL 無毒性量。毒性または毒性影響が観察されなかった最高用量。

OR オッズ比。ばく露と結果との関連性を示す尺度。OR は、特定のばく露を受けた場合に転帰が起こる確率を、ばく露がなかった場合に転帰が起こる確率と比較して表している。

PBTK-TD Physiologically based toxicokinetic/toxicodynamic modelling (PBTK-TD) Physiologically based toxicokinetic/toxicodynamic modelling とは、生理学的プロセスに関する先端的な知識を他の既知/観察された情報と統合して、ヒト、前臨床試験動物種及び／または他の生物の体内での化合物の転帰と影響を模倣することを目的とした数学的モデル化手法である。

PPP 植物防疫製品（農薬）。用語「pesticide（殺虫剤）」はしばしば「plant protection product（植物防疫製品）」と互換的に使用されるが、pesticide（殺虫剤）は植物／作物以外の用途、例えば biocide（殺生物剤）などもカバーするより広い用語（農薬）である。

RR 相対リスク（Relative risk）。ある事象（病気の発生など）がばく露したグループで発生する確率と、比較対照の非ばく露グループで発生する確率との比。

RMS Rapporteur（ラポーター）の加盟国。農薬有効成分の毒性評価に関する書類の評価及び評価を最初に担当する欧州連合の加盟国。

感度 ある検査で個人を正しく「疾病」と同定する能力。疾患が存在する場合に検査が陽性となる可能性。

特異性 個人を疾患無しと正しく同定する検査の能力。疾患がない場合に検査が陰性である可能性。

代替エンドポイント（surrogate endpoint）臨床エンドポイントの代わりとなることを目的としたバイオマーカー。

AHS 農業健康調査

ASHTIII 化学物質による健康影響の脅威に対する警告と報告システム、フェーズ III

BEES-C バイオモニタリング、環境疫学、短命化学物質

DAR 評価報告書草案

DDE ジクロロジフェニルジクロロエチレン

DDT ジクロロジフェニルトリクロロエタン

EMA 欧州医薬品庁

EPA 米国環境保護庁

EQUATOR 健康研究の質と明白性を高める

EU-OSHA 欧州労働安全衛生機関

EWAS Exposome-wide association studies エキスポソームワイド関連研究

GIS 地理情報システム

GLP	優良試験所規範
GPS	全地球測位システム
HWE	健康労働者効果
IATA	試験と評価に関する統合的アプローチ
ICD	国際疾病分類
IHR	国際保健規則
INSERM	フランス国立保健医療研究所
LOQ	定量化限界
MGUS	単クローン性ガンマグロブリン血症
MIE	標的分子への作用
MOA	作用機序
NIHL	非ホジキンリンパ腫
NIOSH	国立労働安全衛生研究所
NOS	ニューカッスルオタワスケール
OECD	経済協力開発機構
OPP	農薬プログラム局
PCC	毒物管理センター
PPE	個人用保護具
RAR	更新評価報告書
RASFF	食品と飼料をカバーする迅速警報システム
RTI	リサーチトライアングル研究所
SAR	構造活性相関
STREGA	遺伝学的関連研究への STROBE 拡張
STROBE	疫学における観察研究の報告の強化
UF	不確実性因子
WHO	世界保健機関
WOE	エビデンスの重み付け

付属書 A－EFSA の外部科学報告書でレビューされた農薬疫学研究及びその他のレビュー

EFSA 外部科学報告書（Ntzani ら、2013 年）によって収集された広範なエビデンスは、疫学研究からの農薬ばく露と健康影響に関するかなりの量の情報が利用可能であることを強調している。それにもかかわらず、このエビデンスの質は通常低く、多くのバイアスが結果に影響を与え、結論を出すことができない可能性が高い。特に、ばく露疫学は長い間、測定と定義の貧弱さに悩まされてきたが、特に農薬に関しては、これは常に評価と定義が非常に難しいものであった。

A.1. EFSA の外部研究報告書

A.1.1. 方法論的品質評価

外部研究報告書は、2006 年 1 月 1 日から 2012 年 9 月 30 日までに発表されたすべての疫学研究の包括的なシステマティックレビューから構成されており、農薬ばく露とヒト健康関連影響の発生との関連性を調査している。

対象となる研究の方法論的評価（各研究に関連するバイアスのリスクを評価する）は、研究デザイン、研究対象集団、ばく露の定義の詳細度、ばく露の測定方法、測定の特殊性に焦点を当てた。マッチングモデルや多変量モデル、盲検化されたばく露評価、十分に説明された有効な結果評価を通じた交絡因子の説明などの取り組みが検討された。

方法論的評価の要素は、Research Triangle Institute（RTI；Research Triangle Park, NC, USA）の項目バンクで検討されたもので、観察による研究の偏りのリスクと精度を評価するための実用的で検証済みのツールである。これらの要素を以下に示す（表 A.1）。

表 A.1：Research Triangle Institute（RTI；Research Triangle Park, NC, USA）の疫学研究の方法論的評価のための項目バンクの要素

質問	高リスク	低リスク
研究デザイン（有望、回顧的、混合、断面的）	回顧的、混合、該当なし	有望
除外基準が明確に記載されている（はい、部分的に、いいえ）	いいえ	はい
著者は電力計算について言及しています（はい、いいえ）		はい
暴露の記述の詳細レベル（高、中、低）	低	高
暴露のロバストな測定（バイオマーカー（有）；小面積生態学的尺度、職種、アンケート（部分的）；大面積生態学的尺度（無）に基づいている。）	いいえ	はい
曝露の尺度は特定のものだったか？ はい；より広範な化学的に関連したグループに基づいて（部分的）、多様な化学的および毒性学的特性の広範なグループに基づいて（いいえ）。	いいえ	はい
グループ間の配分のバランスを図る（層別化、マッチングなど）。	いいえ	はい
潜在的な交絡因子の調整を行った（はい、いくつか、いいえ）。	いいえ	はい
被ばく状態に盲検化された評価者（コホート研究の場合）	いいえ	はい
有効かつ信頼性の高い尺度を用いて評価された結果は、すべての研究参加者に一貫して実施されているか？	いいえ	はい
サンプルサイズ	低	最大
ラフな品質評価	6 以上の回答で高リスク	6 以上の回答で低リスク

結果の定量的な統合は、対象となる結果ごとに 5 件以上の適格な研究があり、発表されたエビデンス間に実質的な異質性がない場合に試みられた。出版バイアスは、10 件以上の研究がメタアナリシスに含まれている場合に、非対称性を視覚的に確認できる漏斗プロットを用いて評価した。

毒性学的データは、外部科学研究報告書ではレビューまたは議論されなかった。

A.1.2. 除外基準

レビューの時点で入手可能な疫学的証拠の総合的に評価するために、EU で禁止されているものを含むすべての種類の農薬を検討した。

除外基準。

- 対照集団のない研究（症例報告、症例シリーズ）及び生態学的研究
- 農薬中毒または偶発的な高用量ばく露
- 影響の推定に関する定量的な情報がない研究
- データの重複を避けるために、追跡期間が異なり、同じ健康影響（アウトカム）を調査している研究については、追跡期間が最も長いもののみを残した。
- 様々な病状の治療に使用される物質の有害影響について言及した研究（例：ワルファリンをベースとした抗凝固薬）。
- 農薬中の溶剤やその他の非有効成分（補助剤など）の研究
- ばく露とばく露のバイオマーカーとの関連性を検討した研究は、健康影響を調べないため対象外とされた。
- 農薬へのばく露を調査した研究／解析：ヒ素、ヘキサクロロシクロヘキサン（HCH）a または b、鉛、ダイオキシン類及びポリ塩化ビフェニル（PCB）を含むダイオキシン様化合物は考慮されなかった。
- ナラティブレビューは除外したが、システマティックレビューやメタアナリシスは除外しなかった。

急性中毒または臨床症例のシリーズである出版物、健康影響とは無関係のバイオモニタリング研究、または動物またはヒトの細胞システムで実施された研究は含まれず、ヒトの健康影響を扱った疫学的研究のみが選ばれた。また、関連性を測定するための定量的データを欠く出版物も除外した。

コホート研究、症例対照研究及び横断研究が含まれた。各研究は、研究デザイン、除外・包含基準の正確な記述、ばく露の記述の詳細度、ばく露の測定の強固性、潜在的な交絡因子の調整、健康影響の評価方法、サンプルサイズ等の 12 の基準を含む方法に基づいて適格性の評価を受けた。これら 12 の基準のうち、3 つの基準はばく露の記述・測定の精度に関連しており、多くの疫学研究が選ばれなかった理由を説明することができるかもしれない。

A.1.3. 結果

全体では、602 の個別の論文が科学的レビューに含まれている。これら 602 の出版物は、6,479 の異なる解析に対応していた。エビデンスの圧倒的多数は後ろ向きまたは横断研究（それぞれ 38%と 32%）であり、前向きな研究は 30%のみであった。ばく露評価は研究によって大きく異なり、全体の 46%が農薬ばく露のバイオマーカーを測定し、さらに 46%が農薬ばく露を推定するためにアンケートを使用していた。研究のほぼ半数（49%）がアメリカを拠点としていた。ほとんどの研究では、農薬への職業性ばく露と健康影響との関連性が調査されていた。農薬に関連する疾患の全領域を対象とした研究はこれまで行われていなかった。報告書では、さまざまなアウトカムを調査している（図 A.1）。最も多いのは、がんのアウトカム（N=164）と子どもの健康に関するアウトカム（N=84）である。

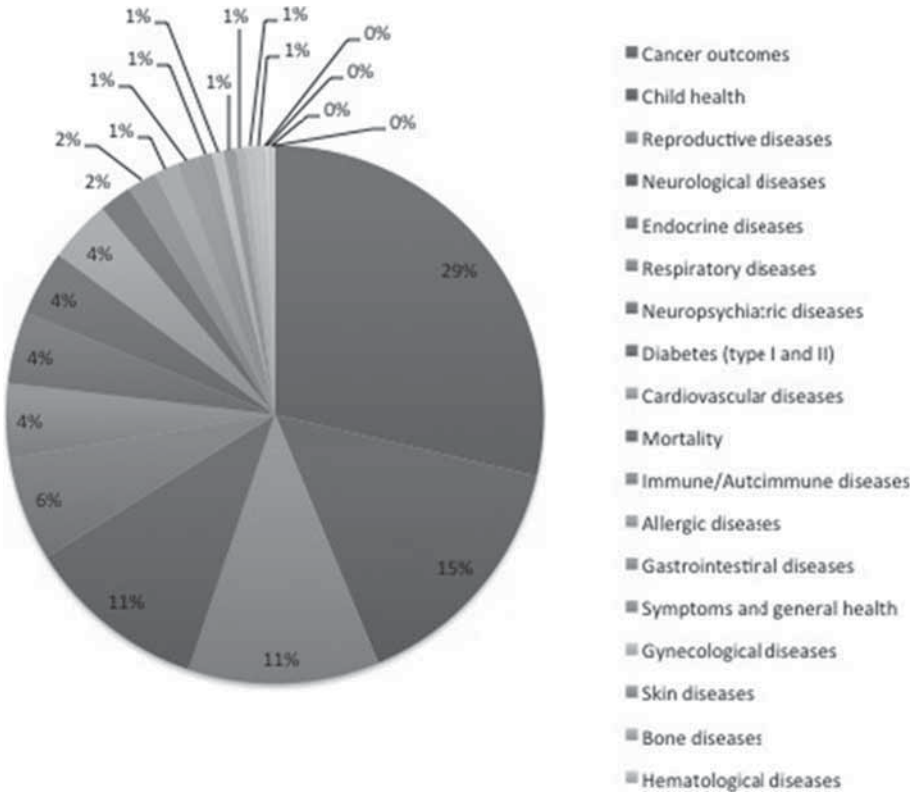


図 A.1：EFSA の外部科学研究報告書でレビューされた出版物のうち、主要な健康影響のカテゴリーと、その健康影響を調査した研究の割合（Ntzani ら、2013 年）

利用可能な大量のデータと大量の分析（>6,000 件）にもかかわらず、研究されたアウトカムの大部分については確固たる結論が出ていない。これは、収集したデータのいくつかの限界と、レビュー自体の固有の限界によるものである。上述したように、レビューでは約 5 年間に農薬に関連して検討されたアウトカムの全範囲が調査された。したがって、最近のエビデンスのみがレビューされ、実施されたメタアナリシスの結果は、利用可能なすべてのエビデンスを含んでいるわけではないため、慎重に解釈されるべきである。したがって、文献全体（5 年以上）をみて、選択したエビデンスの信頼性を評価することに焦点を当てることで、農薬に関連してさらに詳細な解析を行う価値のある結果を浮き彫りにすることができる。研究自体の限界は、環境疫学の他の分野と一致しており、ばく露評価、研究デザイン、統計解析、報告に焦点を当てている。特に以下の点が挙げられる：

a）**ばく露評価**：ばく露の評価は、おそらく ESR でレビューされた研究の中で最も重要な方法論的限界である。研究では、ばく露の評価と割り付けにさまざまな方法が用いられている。ほとんどの研究では、「これまでに使用したことがあるかないか」または「定期的に使用したことがあるかないか」という自己申告による農薬へのばく露に基づいていた。このような方法では、高い誤分類率に悩まされ、用量反応分析を行うことができない。これは特に後ろ向き研究の場合で、病気のある参加者で報告されるより高いばく露量の差が誤分類を生じる（想起バイアス）（Raphael、1987 年）。アンケートは非常に高いばく露レベルと非常に低いばく露レベルの被験者を区別することができるかもしれないが、ばく露濃度による有効な分類を行うことはできず、その結果、用量反応関係の研究を行うことができない。また、ばく露評価のためのアンケートは、疫学研究で使用するために検証される必要がある。それにもかかわらず、多くの研究では検証されていないアンケートを使用しているが、これには内容（アンケートが対象とする有害なばく露源をすべて網羅していない）や基準妥当性（例えば、不正確な想起や質問の誤解）に問題があるかもしれない（Coggon、1995 年）。

調査対象とした農薬のカテゴリーの範囲は広いが、研究では多くの場合、広く定義された農薬のカテゴリーに集中しているため、対象集団がどのような種類の農薬にばく露されているのかを知ることは困難である。

農薬へのばく露は、研究参加者による農薬の使用報告または政府の登録データとして定義された。これらのデータ

は、自己記入式アンケート、質問者が管理するアンケート、職業性ばく露マトリックス（JEM）、居住状況（農薬ばく露に近接しているかどうか）、農薬ばく露に関連するバイオマーカーの検出、または各研究によって定められたその他の方法から得られたものである。

研究では、欧米の集団や EU ですでに禁止されている農薬を調査することが多い。ばく露評価の手段としてバイオマーカーを使用することはまれであるが、ほぼ半数の研究ではまだ利用可能である。

b）**研究デザイン**：上述したように、エビデンスの大部分は症例対照研究と横断研究から得られている。横断研究や一部の症例対照研究では、時間的関係を完全に評価することができないため、関連性の因果関係に関する裏付けを提供することができない。

c）**調査された**健康影響：臨床健康影響の定義は、適格な疫学研究において大きなばらつきを示しており、これが結果のばらつきの原因となっている。おそらくこのような状況において最も重要なのは、調査された多数の代替健康影響の使用である。代替健康影響とは、特定の臨床転帰の代替または予測因子として一般的に受け入れられているバイオマーカーまたは身体測定値のことである。しかし、多くの場合、これらの代替健康影響は検証されておらず、代替健康影響の厳密な定義を満たしていない。このような健康影響は、臨床健康影響の予測因子の可能性があるとされているが、代替健康影響の基準を満たしていない。これらの健康影響の暗黙の仮定を考慮に入れることにより、検証されていない代替健康影響に関するエビデンスを評価することが不可欠である。

広範囲の病態生理をカバーすると評価されたアウトカムは非常に多様だった。データベースに含まれる多くの代替健康影響と同様に「ハード」な臨床健康影響は、評価された臨床研究の課題にアプローチするために支持された異なる方法論を反映している。さまざまな健康影響は 23 の主要な疾患カテゴリーに分類されており、がんと小児の健康影響を扱った研究が最も多くを占めている。

評価された健康への悪影響には以下のものが含まれる。

a）がん、呼吸器（アレルギー）、生殖（受精率低下、先天性疾患）、神経変性（パーキンソン病）などの主要な臨床転帰；

b）神経発達障害（神経認知スケールで評価）などの臨床代替健康影響；

c）検査での代替健康影響（例：肝酵素の変化）。

農薬ばく露に起因する多くの有害な健康影響については、矛盾した、あるいは曖昧な研究が存在する。この結果が一貫性の欠如からくるものなのか、真の不均一性からくるものなのかは、さらなる解明が必要である。

d）**統計的解析**。

異なる供給源からの複数の物質（重金属、溶剤、浮遊粒子状物質など）への同時ばく露は一般的である。それらのすべてが有害な健康影響をもたらす可能性があるため、結果にさらなるバイアスがかかる可能性がある。したがって、真の関連性を明らかにするためには、複数の物質へのばく露による交絡を考慮することが不可欠であるが、EFSA の外部科学研究報告書で評価された圧倒的多数のエビデンスでは、これは不可能であった。

さらに、EFSA の外部科学研究報告書（Ntzani ら、2013 年）で収集・評価されたエビデンスは、都合の良い報告と複数の試験に悩まされる可能性が高い。研究は非常に広範な解析を報告しており、602 の出版物で 6,000 の解析が行われた。多重仮説検定の量は膨大である。これらの解析は、複数の仮説検定のために調整されていなければならず、そうでなければ結果は高い偽陽性率に悩まされる。研究が 1 つの解析しか行われていない場合でも、他の疫学的研究でも示されているように、都合の良い報告の可能性が常にある。さらに、結果を解釈する際には、特に特定の健康影響（がんなど）については、エビデンスの大部分が単一の研究集団そして特に農業健康調査から得られていることも考慮に入れるべきである。

A.1.4. EFSA 外部科学研究報告書の結論

上記で強調された限界にもかかわらず、外部科学研究報告書（Ntzani ら、2013 年）では農薬ばく露とパーキンソン病及び小児の白血病との関連、これらは先行するメタアナリシスでも裏付けられている、について首尾一貫したエビデンスを示した。さらに、肝臓がん、乳がん、II 型糖尿病など、これまであまり研究されてこなかった多様な健康影響につ

いてもリスクの増加が認められた。内分泌疾患、喘息、アレルギー、糖尿病、肥満などの他の健康影響への影響は、リスクの増加を示しており、今後さらに調査が必要である。

小児白血病とパーキンソン病は、2006 年以降のメタアナリシスで一貫して農薬ばく露に関連したリスクの増加を示した 2 つの健康影響である。それにもかかわらず、特定の農薬クラスや個々の農薬の影響を解きほどくためには、ばく露をよりよく研究する必要がある。他の健康影響についても、有意な要約推定値が報告されている(表 A.2 に要約)。しかし、これらは 2006 年以降の研究であるため、結果は関連性を示唆するものとみなすべきであり、特にばく露の不均一性に関する限界は常に考慮されるべきである。存在しうるバイアスの量を定量化し、バイアスの推定値を考慮に入れても農薬との関連性が十分に支持される結果を分離するために、2006 年以前の出版物を含むように結果を更新した後、特定の健康影響に関連してこれらのデータにデータ統合と統計ツールを適用すべきである。同様に、結論を出すためにさらなるエビデンスが必要な健康影響は、強調される必要がある。

表 A.2:報告書で実施されたメタアナリシスの概要

健康面での成果	N studies	メタアナリシスの結果	I^2
白血病	6	1.26 (0.93; 1.71)	59.40%
ホジキンリンパ腫	7	1.29 (0.81–2.06)	81.60%
小児白血病（妊娠中の農薬ばく露）	6	1.67 (1.25–2.23)	81.20%
小児白血病（妊娠中の殺虫剤ばく露）	5	1.55 (1.14–2.11)	65%
小児白血病（妊娠中に殺虫剤にばく露された場合-ターナー，2010 年の更新）	9	1.69 (1.35–2.11)	49.80%
小児白血病（妊娠中に不特定多数の農薬にばく露された場合	5	2.00 (1.73–2.30)	39.60%
小児白血病(妊娠中の特定されていない農薬へのばく露 - ターナー、2010 年の更新)	11	1.30 (1.06–1.26)	26.50%
小児白血病（小児期の農薬ばく露）	7	1.27 (0.96–1.69)	61.10%
小児白血病（小児期の殺虫剤ばく露-小児期の殺虫剤ばく露，更新 ターナー，2010 年の更新）	8	1.51 (1.28–1.78)	0%
小児白血病(小児期に特定されていない農薬へのばく露 - ターナー、2010 年の更新)	11	1.36 (1.19–1.55)	0%
乳がん（DDE ばく露）	5	1.13 (0.81–1.57)	0%
乳がん	11	1.24 (1.08–1.43)	0%
精巣がん（DDE ばく露）	5	1.40 (0.82–2.39)	59.50%
胃がん	6	1.79 (1.30–2.47)	0%
肝臓がん	5	2.50 (1.57–3.98)	25.40%
停留精巣	8	1.19 (0.96–1.49)	23.90%
停留精巣（DDT ばく露）	4	1.47 (0.98–2.20)	51%
尿道下裂（一般的な農薬ばく露）	6	1.01 (0.74–1.39)	71.50%
尿道下裂（特定の農薬へのばく露）	9	1.00 (0.84–1.18)	65.90%
流産	6	1.52 (1.09–2.13)	63.10%
パーキンソン病	26	1.49 (1.28–1.73)	54.60%
パーキンソン病（DDT ばく露）	5	1.01 (0.78–1.30)	0%
パーキンソン病（パラコートばく露）	9	1.32 (1.09–1.60)	34.10%
筋萎縮性側索硬化症	6	1.58 (1.31–1.90)	10%
喘息（DDT ばく露）	5	1.29 (1.14–1.45)	0%
喘息（パラコートばく露）	6	1.40 (0.95–2.06)	53.30%
喘息（クロルピリホスばく露）	5	1.03 (0.82–1.28)	0%
1 型糖尿病（DDE ばく露）	8	1.89 (1.25–2.86)	49%
1 型糖尿病（DDT ばく露）	6	1.76 (1.20–2.59)	76.30%
2 型糖尿病（DDE ばく露）	4	1.29 (1.13–1.48)	0%

N=メタアナリシスのために検討された研究の数；メタアナリシス結果の列では、数字は効果の大きさ（オッズ比（OR）または相対リスク（RR））の統計的推定値を、対応する 95%信頼区間（CI）とともに表している。

I^2 は、研究間の総変動のうち、不均一性に起因する割合を示す。

A.2. INSERIM レポート

2013 年 9 月、フランス国立保健医療研究所（INSERM）は、農薬へのばく露によるヒト健康影響について専門家グループと共に実施した文献レビューを発表した[22]。2012 年 6 月までの科学文献に発表された疫学的または実験的データが解析された。報告書には、文献的解析を概説し、主要な結論と政策及び勧告を強調した要約が添付されている。

INSERM の報告書は 4 つの部分から構成されている。(1) ばく露評価、疫学研究におけるばく露を評価するための直接的及び間接的な方法の詳細な説明、(2) 疫学、2012 年までの文献で利用可能な疫学研究のインベントリと解析及び推定される関連性の強さを評価するためのスコアリングシステム、(3) 毒性学、いくつかの物質の毒性学的データ（代謝、作用機序、分子経路）のレビューと生物学的妥当性の評価、(4) 推奨事項。

健康影響の発生と推定される中等度または強い関連性を持つと INSERM の報告書で特定された物質の大部分は、現在禁止されている化学物質である。これは主に、調査された疾患の大部分が高齢者の疾患であるという事実によって推進されている。したがって、これまでに実施された研究は、研究の時点で高齢であり、何年も前にばく露された人に基づいている。結論から言うと、最近の製品の多くの潜在的な長期的影響を調査することはまだ可能ではない。

これらの物質は、DDT やトキサフェンのような有機塩素系殺虫剤や、テルブフォスやプロポキサーのようなコリンエステラーゼ阻害作用を持つ殺虫剤のグループに属している。

INSERM の専門家評価報告書で確認された 7 つの承認された有効成分（除草剤 2,4-D、MCPA、メコプロップ、グリホサート、殺虫剤クロルピリホス、葉の殺菌剤マンコゼブとマネブ）のうち、すべてが造血器がんとの中等度または弱い関連性があると推定されていた。そのうち 2 つ（葉状殺菌剤マンコゼブとマネブ）はパーキンソン病との関連性が弱いと推定され、2 つ（クロルピリホスとグリホサート）は専門家の評価で弱い、または中等度とされた発達障害との関連性があると推定された。

A.2.1. 疫学研究におけるばく露評価方法の説明

ばく露を評価するために、生物学的または環境モニタリングデータ、その場しのぎのアンケート、職業別または作物別のばく露マトリクス、専門家のカレンダーの分析、販売データ、土地利用データなど、さまざまな方法（直接的及び間接的）が開発されてきた。著者らによると、これらの様々なツールは互いに組み合わせることができるが、現在までのところ、職業性農薬ばく露評価の背景でばく露を推定するための基準となる方法として有効性が確認されていない。

A.2.2. 疫学

INSERM の専門家グループは、文献で入手可能な疫学研究の目録作成と解析を行い、農薬ばく露と健康影響との関連性の可能性を検討した。8 つのがん部位（非ホジキンリンパ腫、白血病、リンパ腫、多発性骨髄腫、前立腺、精巣、脳、メラノーマ）、3 つの神経変性疾患（パーキンソン病、アルツハイマー病、筋萎縮性側索硬化症）、認知・抑うつ障害、生殖能への影響（受胎性、胎児と出生児の発生）、小児がんである。これらは、以前の研究で農薬ばく露に関連する可能性があるとして同定されている健康影響である。

主に農家、農薬散布者、農薬製造業の労働者を対象とした疫学的研究と、関連性がある場合には一般集団を対象とした研究が選ばれた。

INSERM の専門家グループは、研究の関連性における階層を設定し、メタアナリシスを最上位に置き、次にシステマティックレビュー、コホート研究、そして最終的には症例対照研究とした。この階層に基づいて、研究結果の解析から、ばく露と健康影響の発生との間の関連性の推定の強さを評価するためにスコアリングシステムが定義された；調査された各疾患や病態について、このスコアは、例えば、利用可能な研究の質、種類、数によって異なる。

（++）：強い推定：メタアナリシスの結果に基づく、または複数のコホート研究、または少なくとも 1 つのコホート研究と 2 つの症例対照研究、または 2 つ以上の症例対照研究。

（+）：中程度の推定：コホート研究または入れ子になった症例対照研究または 2 つの症例対照研究の結果に基づ

[22] INSERM. 農薬。sante への影響。Collection expertise collective, Inserm, Paris, 2013.

く。

(±):弱い推定:1 件の症例対照研究の結果に基づく。この統合により、この作業は単純なマッピング作業の状態を超えたものとなった。

A.2.3. 毒性学的データ

文献レビューで検討した毒性データは、主に代謝、作用機序、分子経路に関するものであった。製品を上市するための手続きの一部として提供された研究は、公表された文献で発表された場合を除き、考慮されなかった。

疫学研究で物質が明らかに同定された場合には、研究結果から生物学的に妥当であるかどうかを評価するためのスコアリングシステムを設定した:病態生理学的データとの整合性と健康影響の発生。

(++):3 つの毒性メカニズムで支持された仮説。

(+):少なくとも 1 つの毒性メカニズムによって支持された仮説。

A.2.4. 所見

INSERM 報告書の主な結果は、表 A.3-A.6 にまとめられている。

表 A.3:農薬への職業上ばく露と成人の健康影響との間の統計的に有意な関連(レビューで解析された健康影響)

健康面での成果	リスク過剰が顕著な集団のタイプ	推定の強さ (a)
エヌエイチエル	農家、農薬散布者、製造工場関係者	++
前立腺がん	農家、農薬散布者、製造工場関係者	++
多発性骨髄腫	農家、農薬散布者	++
パーキンソン病	職業上ばく露と非職業上ばく露	++
白血病	農家、農薬散布者、製造工場関係者	+
アルツハイマー病	農家	+
認知障害 (b)	農家	+
妊孕性・胎動性障害	職業上ばく露	+
ホジキンリンパ腫	農業従事者	±
精巣がん	農業従事者	±
脳腫瘍(神経膠腫、髄膜腫	農業従事者	±
メラノーマ	農業従事者	±
筋萎縮性側索硬化症	農家	±
不安、うつ病 (b)	農家、急性中毒の既往歴のある農家、農薬散布者	±

(a):スコアリングシステム:強い推定(++)、中程度の推定(+)、弱い推定(±)。
(b):ほとんどの農薬が有機リン酸塩であった。

表 A.4:職業または家庭生活での農薬ばく露と小児のがんまたは発達障害(レビューで解析された健康影響)との間の関連(統計的に有意な関連のみを示す)

健康面での成果	リスク超過が顕著なばく露の種類と集団	推定の強さ(a)
白血病	妊娠中の職業上ばく露、出生前ばく露（住居）	++
脳腫瘍	妊娠中の職業上ばく露	++
先天性奇形	妊娠中の職業上ばく露。	++
	妊娠中の住居ばく露（農地、家庭生活での使用）	+
胎児の死	妊娠中の職業上ばく露	+
神経発達	妊娠中の住居ばく露（農地、家庭生活、食品）(b)。	++
	妊娠中の職業上ばく露	±

(a)：スコアリングシステム：強い推定（＋＋)、中程度の推定（+)、弱い推定（±)。
(b)：有機リン酸系。

表 A.5:承認された有効成分に関する知見:疫学的評価と生物学的妥当性

有効成分	分類	推定の強さ (a)	生物学的妥当性 (b)
有機リン酸系殺虫剤			
クロルピリホス	急性毒性 CAT3	白血病（+) 神経発達（+) ＮＨＬ（±)	Yes（++) Yes（++) Yes（++)
ジチオカルバメート系殺菌剤			
マンコゼブ／マネブ	生殖毒性 CAT2	白血病（+) メラノーマ（+) パーキンソン病（パラコートと併用して）(±)	? ? Yes（+)
フェノキシ系除草剤			
2,4-D	急性毒性 CAT4	NHL（+)	?
MCPA	急性毒性 CAT4	ＮＨＬ（±)	?
メコプロップ	急性毒性 CAT4	ＮＨＬ（±)	?
アミノホスホン酸グリシン除草剤			
グリホサート		NHL（+) 胎児死亡（±)	? ?

(a)：スコアリングシステム：強い推定（＋＋)、中程度の推定（+)、弱い推定（±)。
(b)：スコアリングシステム。(++)：毒性の 3 つの異なる既知のメカニズムによって支持された仮説、(+)：毒性の少なくとも 1 つのメカニズムによって支持された仮説。

表 A.6:非承認有効成分に関する知見:疫学的評価と生物学的妥当性

有効成分	Ban in the EU	IARC classfication	推定の強さ[a]	生物学的妥当性[b]
ディルドリン	1978	3or2（US-EPA）	ＮＨＬ（c） 前立腺がん（±） パーキンソン病（±）	Yes（+） Yes（+） ?
DDT／DDE	1978	2B	NHL（++） 精巣がん(+) 子供の成長（++） 神経発達（±） 精子パラメータ異常（+）	Yes（+） ? ? ? ?
クロルデン	1978	2B	ＮＨＬ 白血病（+） 前立腺がん（±） 精巣がん（+）	Yes（+） Yes（+） Yes（+） ?
リンデン（c-HCH	2002/2004/2006/2007	2B(d)	NHL（++） 白血病（+）	Yes（++） Yes（++）
b-ＨＣＨ	2002/2004/2006/2007	2B(d)	前立腺がん（±）	?
トキサフェン	2004	2B	ＮＨＬ（c） 白血病（+） メラノーマ（+）	Yes（++） Yes（++） Yes（+）
クロルデコン	2004	2B	前立腺がん（++） 精子パラメータ異常（+） 神経発達（+）	Yes（+） ? ?
ヘプタクロル	1978	2B	白血病（+）	Yes（+）
エンドスルファン	2005	Not classified	?	Yes（+）
ヘキサクロロベンゼン（ＨＣＢ）	1978	2B	子供の成長（+）	?
テルブフォス	2003/2007		NHL（+） 白血病（+）	? ?
ダイアジノン	2008		NHL（+） 白血病（+）	? ?
マラチオン	2008	3	NHL（++） 白血病（+） 神経発達（+） 精子パラメータ異常（+）	Yes（+） Yes（+） ? ?
フォノフォス	2003		ＮＨＬ（±） 白血病（+） 前立腺がん（+）	? ? ?
パラチオン	2002	3	メラノーマ（+）	?
クマフォス	EU では届出認可されていない		前立腺がん（+）	?
カルバリル	2008	3	ＮＨＬ（±） メラノーマ（+） 精子パラメータ異常（+）	? ? ?
プロポキサー	2002		神経発達（+） 胎児発育（+）	? ?
カルボフラン	2008		ＮＨＬ（±） 前立腺がん（+）	? ?
ブチル酸塩	2003		NHL（+） 前立腺がん（+）	? ?
EPTC	2003		白血病（+）	?
アトラジン	2005	3	ＮＨＬ 胎児発育（+）	Yes（+） ?
シアニジン	2002/2007		ＮＨＬ（c）	?
ペルメトリン	2002	3	前立腺がん（+）	Yes（+）

フェンバレレート	1998	Not classified	精子パラメータ異常（+）	?
臭化メチル	2010	3	精巣がん（+）	?
ジブロモエタン	Banned	2A	精子パラメータ異常（+）	?
ジブロモクロロプロパン（DBCP）	Banned	2B	精子パラメータ異常／不妊（＋＋＋＋）（因果関係）	Yes（+++）（作用機序解明）
パライバット	2007		パーキンソン病（+）	Yes（++）
ロテノン	2011		パーキンソン病（+）	Yes（++）
アラクロール	2008		白血病（+）	Yes（++）

（a）：スコアリングシステム：強い推定（＋＋）、中程度の推定（+）、弱い推定（±）。
（b）：スコアリングシステム。（++）：3 つの毒性メカニズムに支持された仮説、（+）：少なくとも 1 つの毒性メカニズムに支持された仮説。
（c）：t（14,18）転座を有する母集団のみ。（d）：技術的混合物（α-、β-及び γ-ＨＣＨ）。

A.2.5. 推奨事項

いくつかの有効成分に関する利用可能な疫学的・機序学的データを解析した結果、さらなる研究開発のためのいくつかの推奨事項が示唆された。

a）農薬への集団ばく露に関する知識は改善されるべきである。

1）農家の有効成分使用に関する情報収集
2）実際のばく露レベルを測定するための農地での研究の実施
3）生涯労働期間のばく露を監視すること
4）空気（屋外・屋内）、水、食品、土壌中のばく露レベルの測定
5）急性中毒に関する情報収集
6）バイオモニタリングや外部ばく露量測定のための解析方法の改善
7）研究者が広範な製剤データ（溶剤、共配合製剤など）にアクセスできるようにする。

b）ばく露と健康影響との間の潜在的な関連性を研究する。

1）健康影響をもたらす物質または物質群の特性を把握する
2）影響を受けやすい個人または集団に焦点を当てる（酵素の遺伝子多型など）
3）ばく露枠と感受性（妊娠期間、発育期間）に焦点を当てた研究
4）疫学と毒物学のギャップを埋める（作用機序）
5）混合物の毒性に関する知識の向上
6）研究の新しいアプローチ（in vitro や in silico モデル、オミクスなど）を育成する。

A.3. EFSA 外部科学研究報告書と INSERM 報告書の類似点と相違点

ここで議論されている 2 つの報告書は、異なる方法論を使用している。しかし、多くの場合、それらの結果と結論は一致している。INSERM の報告書は、事前に調査された結果に限定されており、毒物学的データもレビューすることで疫学研究の生物学的妥当性を調査しようとしているのに対し、EFSA の報告書は、約 5 年の期間に発表されたすべての利用可能な疫学研究の包括的なシステマティックレビューである。

両報告書の違いは表 A.7 に示されており、検索期間（すなわち、両報告書は同じ出版データを評価していない）、研究の適格性の基準の違い、健康影響全体と健康影響内のエビデンスを要約するアプローチの違いに関連している。

全体的に、INSERM の報告書は EFSA の報告書よりも多くの健康被害との関連を特定している。しかし、同じ健康影響（小児白血病、パーキンソン病）については、両方の報告書で農薬ばく露との関連性が十分に証明されていると主張されている。

表 A.7:EFSA 外部科学報告書と INSERM 報告書で使用された方法の比較

	EFSA External report	INSERM report
レビューされた論文	602/43,000	NR
言語	Yes	NR
検索戦略（キーワード、MeSH）	Yes	NR
検索データベース	Yes (4)	NR
出版年	2006–2012 (Sep)	? to 2012 (Jun)
評価された疫学研究の種類	Cross-sectional Case–control Cohort	Cross-sectional Case–control Cohort
含有基準	Yes	NR
除外基準	Yes	NR
方法論的品質評価	Yes (12 criteria)	NR
ばく露グループ(a)	Yes	Yes
ばく露評価	Yes	Yes
定量的統合（メタアナリシス）	Yes	No
質的統合(c)	Yes	Yes
毒物学的データのサポート	NI	Yes
個々の農薬との関連性	Yes	Yes
健康影響研究		
血管がん	Yes	Yes
充実性腫瘍	Yes	Yes
小児がん	Yes	Yes
神経変性疾患	Yes	Yes
神経発達影響	Yes	Yes
精神障害(b)	No	Yes
生殖と発生	Yes	Yes
内分泌	Yes	NI
代謝	Yes	Yes
免疫学的	Yes	NI
呼吸器	Yes	NI

NR：報告されていない、NI：調査されていない。
(a)：ばく露の種類（環境、職業など）及び期間（一般集団、児童など）。
(b)：例：うつ病性障害。
(c)：説明を追加する。

A.4. The Ontario College of Family Physicians の文献レビュー（OCFPLR）

2004 年、カナダの The Ontario College of Family Physicians は、1992 年から 2003 年の間に発表された、農薬ばく露に関連した主要な健康影響に関する文献をレビューした。著者らは、表 A.8 に示すように充実性腫瘍と農薬ばく露との間には正の関連が存在すると結論づけた。著者らは、よく計画された大規模コホート研究では、これらの関連性は一貫して統計的に有意であり、その関係は高ばく露レベルで最も一貫していたと指摘している。また、用量反応関係がしばしば観察され、研究の質は概ね良好であるとした。

表 A.8:オンタリオ州家庭医大学のレビューで考慮された健康影響、2004 年

エンドポイント	オンタリオ大学で同定された関連性、農薬（差異化されている場合）、研究の種類、（研究数／総研究数）
A）がん	
1. 肺	—ve cohort（1/1） +ve case–control（1/1） +ve carbamate, phenoxy acid, case–control（1/1）
2. 乳房	+ve case–control（2/4） +ve ecological（1/1） +ve triazine, ecological（1/1） —ve atrazine, ecological（1/1）
3. 大腸直腸	
4. 膵臓	+ve cohort（1/1） +ve case–control（2/2）
5. 非ホジキンリンパ腫	+ve cohort（9/11） +ve case–control（12/14） +ve ecological（2/2）
6. 白血病	+ve cohort（5/6） +ve case–control（8/8） —ve ecological（1/1） +ve lab study（1/1）
7. 脳	+ve cohort（5）, similar case–control（5）
8. 前立腺	+ve cohort（5/5）case–control（2/2）ecological（1/1）
9. 胃	
10. 卵巣	
11. 腎臓	+ve pentachlorophenol cohort（1/1） +ve cohort（1/1） +ve case–control（4/4）
12. 精巣	
B）非がん	
1）生殖影響	+ve glyphosate
先天性奇形	+ve pyridyl derivatives
分娩/妊娠までの期間	Suggest impaired
受胎性	
発育影響	Possible +ve association, but further study required
胎児死亡	Suggested association
混合した健康影響	
2）遺伝毒性・免疫毒性 染色体異常 ＮＨＬ再編成	+ve Synthetic pyrethroids（1） +ve organophosphates（1） +ve fumigant and insecticide applicators +ve fumigant and herbicide applicators
3）皮膚科学的	
4）神経毒性の精神及び感情への影響	+ve
機能的神経系の影響	+ ve organophosphate/carbamate poisoning
神経変性の影響（PD）	+ve cohort（4/4） +ve case–control（2/2） +ve ecological（1/1）

+ve: positive; —ve: negative.

報告書は、農薬ばく露と非ホジキンリンパ腫（NHL）の発症との関連性を示す説得力のあるエビデンスがあり、また農薬ばく露と白血病との間に正の関連性があるという明確なエビデンスがあると結論づけている。著者らはまた、様々なばく露時間の経過から生じる多くの神経系への影響についても一貫した結果が得られたと主張している。

このような断定された結論は、非政府組織（NGO）の支持を得て、いくつかの規制当局の間で疑問が生じた。当時の

英国政府の独立諮問委員会である農薬諮問委員会（ACP）は、オンタリオ大学レビューの結果の評価を依頼された。委員会のメンバーには 1 人の疫学者が含まれており、委員会は、他の政府の委員会に独立した助言を提供することに関与している他の 5 人の疫学者に相談した。彼らはすべてのレビューが主要な欠点を持っていたことに同意した（例えば、正確な検索戦略と特定されていない選択基準、結果の選択的な報告、不適切な理解と関連する毒性学の考慮、ばく露のルートとレベルへの不十分な注意、正当化された結論、等）。全体的に、オンタリオ大学レビューの結論は、提示された解析によってサポートされていないと考えられた。2012 年には、オンタリオ大学レビューの著者は、彼らの評価の更新を発表した；彼らの 2 番目の報告書では、彼らは非常に似たようなアプローチを使用したが、使用される包含基準に関するより詳細を提供した。この例は、疫学研究の過剰解釈のリスクを思い起こさせるものである。特に、ばく露と有害な性健康影響の発生との間の因果関係を推論することはよくあるが、これはさらに評価されるべき関連性を示している。

付属書 B－EFSA が委託したヒト・バイオモニタリング・プロジェクト[23]

2015 年、EFSA は、疫学研究におけるばく露評価のためのツールとして、また、農薬への職業上ばく露による潜在的な健康リスク評価に貢献するために、労働安全衛生戦略における HBM の役割をさらに調査するためのプロジェクトを外部委託した。実際、ばく露評価はすべての疫学研究の重要な部分であり、ばく露の誤分類や単純な分類法の使用は、接触と健康被害の結果との間に関連性があるかどうかを判断する研究の能力を弱めてしまうことが知られている。

Risk & Policy Analysts Limited（RPA）、IEH Consulting Limited（IEH）、Health & Safety Laboratory（HSL）からなるコンソーシアムは、1990 年から 2015 年までの期間、系統的な文献レビューを実施した。その目的は、作業ばく露評価の再開発のためのツールとしての HBM の使用に関する概要を提供し、長所、短所、さらなる開発の必要性を特定することであった（第一の目的）。検索では、農薬（またはその代謝物）への職業上ばく露を評価するための HBM の使用に関連する 2096 の文献を特定した。検索の結果（Bevan ら、2017 年）は、過去 10～20 年の間に HBM の使用が拡大してきたこと、特に環境や消費者のばく露分析の分野にまで拡大してきたことを示している。しかし、農薬ばく露評価のための HBM の使用については、特に、分析品質の向上や標準化のための戦略の開発、代謝物のための標準物質の利用可能性の向上、数学的モデリングへの HBM データの統合、ばく露の再構築、分析機器の改善、ヒト毒性データの利用可能性の向上など、さらなる改善が必要とされている。

請負業者は、EU／米国の作業環境で実施された利用可能な HBM 研究／サーベイランスプログラムのレビューを実施し、残留性のあるものと残留性のないものの両方の農薬（または代謝物）を特定した。最も関連性の高い研究を特定するために、HBM、疫学的、毒性学的側面の品質スコアリングを含む 2 段階のスクリーニングプロセスを利用し、178 件の研究をクリティカルレビューの対象とした。特定された研究のスクリーニングと並行して、これらの文献からのデータを照合するためにマスタースプレッドシートが計画され、その中には、研究タイプ、研究参加者、調査対象の化学物質、バイオマーカーの品質チェック、分析方法、ばく露評価、健康影響／毒性エンドポイント、追跡期間、結果の説明、バイアスのリスク、その他のコメントに関する情報が含まれている。

HBM は、様々な農薬への作業者のばく露を監視するために広く使用されている。職業上の農薬の使用に関する疫学的研究では、不十分なばく露情報や後ろ向きなばく露情報が制限されていることが見受けられる。職業別または作物別のばく露マトリックスの使用例も報告されている。しかし、実際のばく露データに対するこれらのマトリックス研究の検証はほとんど行われていない。季節的なばく露と PPE の影響を調査したデータは非常に限られており、多くの研究では、1 つまたは 2 つの特定化合物のみを評価するために HBM を使用している。現在、健康リスク評価には多種多様なばく露モデルが採用されており、モデルによって予測されたばく露推定値を評価するためにバイオマーカーもしばしば使用されている。

関連性があると判断された 178 の出版物から、41 の個別研究が除草剤を含み、そのうち 34 の個別除草剤が同定され、そのうちの 15 が現在 EU での使用が承認されている。同様に、殺虫剤を含む 90 件の個別研究のうち、79 件の殺虫剤が同定され、そのうち 18 件は現在 EU での使用が承認されている。20 の個別研究には殺菌剤が含まれており、34 種類の殺菌剤が確認され、そのうち 22 種類が現在 EU での使用が承認されている。最も研究された除草剤は（順に）、2,4-D>アトラジン>メトラクロール=MCPA>アラクロール=グリホサートであることが示されている。同様に、最も研究された殺虫剤（順に）は、クロルピリホス>ペルメトリン>シペルメトリン=デルタメトリン>マラチオンであり、最も研究された殺菌剤は、キャプタン>マンコゼブ>フォルペットであった。

現在の限界は、特にヒトを対象とした個々の農薬の ADME に関して、ヒトからの動特性データの数が限られていることに起因しており、これにより、すべてのばく露経路についてより正確な HBM サンプリングが可能になる。このことは、毒物動態データに依存する農薬のリスク評価のための PBPK モデルの開発や、現在使用されているばく露評価モデルのバリデーションにも影響を及ぼす。現在、この分野での HBM の使用に影響を与えているさらなる限界は、現在使用されている農薬への長期ばく露を評価するための大規模な前向きコホート研究が不足していることである。

特定されたエビデンスは、ヨーロッパにおける農薬の労働衛生サーベイランスの一環としての HBM の実施に関する

[23] Bevan ら

推奨事項を策定するために使用されている。実施を可能にするために克服しなければならないいくつかの重要な問題が検討された。その中には、新しいスペックや感度の高いバイオマーカーの開発の優先順位の設定、健康に基づいたガイダンス値の導入と採用、研究間の計測値を検証するための QA スキームの開発、野外作業やアンケートの作成における良好な実施、バイオバンキングの使用範囲の拡大、農薬の安全性の承認後のモニタリングにおける HBM の使用などが含まれている。

付属書 C－ハザードの特定のための疫学研究の統合に関する国際規制機関の経験

C.1. WHO-国際がん研究機関（IARC）

国際がん研究機関（IARC）の「ヒトに対する発がん性リスク評価に関する IARC モノグラフ」は、ヒトのがんリスクを増加させる可能性のある環境ばく露を評価するために 40 年前に設立されたプログラムである。これらには、個々の化学物質や化学物質の混合物、職業上ばく露、物理的要因、生物学的要因、生活様式の要因が含まれる。

IARC は、科学者からなる国際的な学際的作業部会を組織し、科学的出版物からのエビデンスの質と強度をレビューして評価し、懸念される物質がヒトに発がんリスクをもたらす可能性を評価するためのハザード評価を実施している。特に、IARC ワーキンググループのメンバーの役割は、がんに関する疫学的研究やその他の実験的研究の結果の評価、発がんのメカニズムに関するデータの評価、ヒトへのばく露による発がん性の総合的な評価を行うことである。

モノグラフは、世界中の政府、組織、公衆衛生の予防・管理措置を設定するために広く利用され、参照されている。

IARC モノグラフの前文[24]は、プログラムの範囲、モノグラフの開発に使用される科学的原理と手順、考慮されるエビデンスの種類、評価の指針となる科学的基準を説明している。モノグラフの範囲は、単一の化学物質だけでなく、関連する化学物質のグループ、複雑な混合物、職業上ばく露、物理的・生物学的物質、生活様式の要因を含むように拡大された。そのため、モノグラフのタイトルは「ヒトに対する発がん性リスクの評価」となっている。

関連する疫学研究、実験動物を用いた発がんバイオアッセイ、メカニズムデータ、ばく露データなどが批判的にレビューされている。公表されている科学文献に掲載された、または掲載が認められた報告書のみが含まれる。しかし、研究を含めることは、研究デザインの妥当性や結果の分析と解釈を受け入れることを意味するものではない。利用可能な研究の質的側面は慎重に精査されている。

モノグラフではハザードの特定を強調しているが、がんのハザードを評価するために用いられた疫学研究や実験研究と同じものを、用量反応関係を推定するためにも用いることができる。モノグラフは、利用可能な疫学データの範囲内で用量反応関係を推定することもあれば、実験研究と疫学研究の用量反応情報を比較することもある。

モノグラフの構成は以下のようになっている。

1）ばく露データ
2）ヒトにおけるがんの研究
3）実験動物を用いたがんの研究
4）メカニズム等の関連データ
5）まとめ
6）評価と根拠

ヒトの疫学的データは、関連するすべての疫学的研究が評価されている 2）に記載されている。バイオマーカーの研究は、ヒトに対する発がん性の評価に関連する場合に含まれる。

疫学研究の IARC 評価には、以下の基準の評価が含まれる：検討された研究の種類（例：コホート研究、症例対照研究、相関（または生態学的）研究及び介入研究、症例報告）、研究の質（例：バイアス、交絡、生物学的変動及び影響の推定精度に対するサンプルサイズの影響）、メタアナリシス及びプール分析、時間的影響。例えば、最初のばく露時の年齢、最初のばく露からの時間、ばく露の期間、累積ばく露、ピークばく露などの時間的変数）、疫学研究におけるバイオマーカーの使用（例えば、ばく露のエビデンス、初期効果のエビデンス、細胞、組織または生物の反応のエビデンス）、因果関係の基準。

因果関係に関する特定の基準では、問題の薬剤がヒトに対して発がん性があるというエビデンスの強さに関して判断がなされる。

判断を下す際に、作業部会は因果関係についていくつかの基準を考慮している（Hill、1965 年）。強い関連性（例えば、大きな相対リスク）は因果関係を示す可能性が高い。しかし、疾患やばく露が一般的な場合には、弱い関連性が重要であることが認識されている。異なるばく露条件で計画の異なる複数の研究で再現された関連性は、単一の研究

[24] http://monographs.iarc.fr/ENG/Preamble/CurrentPreamble.pdf

からの孤立した観察よりも因果関係を示す可能性が高い。異なる研究間で一貫性のない結果が得られた場合には、考えられる理由（例えば、ばく露の違い）を探り、方法論的に健全でない研究よりも質の高い研究の方が重要視される。明確な用量反応効果がないからといって必ずしも因果関係を否定するエビデンスにはならないが、ばく露に伴ってリスクが増加することは因果関係を示す強いエビデンスと考えられている。ばく露の停止または減少後にリスクが低下していることが示された場合も、結果の因果関係の解釈を支持するものである。時間性、影響の推定精度、生物学的妥当性、全体的なデータの一貫性が考慮される。バイオマーカー情報は、疫学的観察の生物学的妥当性の評価に使用されることがある。被験者と非被験者でがんの発生率が異なることを示している無作為試験は、特に因果関係を示す強力なエビデンスとなる。

疫学的研究でばく露とがんとの間の関連性がほとんど、または全く示されない場合には、発がん性がないと判断することができる。このような場合、研究は、バイアスの可能性、交絡の可能性、またはばく露の誤分類を含めて、上述の計画と解析の基準を評価するために精査される。さらに、方法論的に健全な研究は、観察されたばく露レベルのどのようなばく露についても、影響の推定値が一致していること、相対リスクのプール推定値がほぼ一致すること、そして狭い信頼性間隔を持つことに一貫性があるべきである。さらに、個々の研究も、すべての研究のプール結果も、ばく露レベルの増加に伴うリスクの増加を示すべきではない。発がん性がないというエビデンスは、研究されたがんの種類、報告されたばく露量及びこれらの研究で観察された最初のばく露と疾患発症の間の期間にのみ適用できる。ヒトのがんの経験から、最初のばく露から臨床症状のがんの発生までの期間が 20 年よりも長いことがあり、30 年よりも実質的に短い潜伏期間は、発がん性の欠如のエビデンスを提供できないことが示されている。

最後に、疫学研究の結果、標的臓器または組織、用量反応関連、ヒト及び動物のデータのエビデンスの強固さの評価、メカニズムのエビデンスの強固さをまとめた総合評価に到達するために、エビデンスの全体像を検討する。

総合評価の最後に、以下のいずれかのグループに分類される。グループ 1、その薬剤はヒトに対して発がん性がある;グループ 2A、その薬剤はおそらくヒトに対して発がん性がある;グループ 2B、その薬剤はヒトに対する発がん性が疑われる;グループ 3、その薬剤はヒトに対する発がん性に関して分類できない;グループ 4、その薬剤はおそらくヒトに対して発がん性がない。

薬剤の分類は、ヒト及び実験動物での研究、ならびにメカニズム及びその他の関連データから得られたエビデンスの強固さを反映する科学的判断の問題である。これらの分類は、ばく露が発がん性であるというエビデンスの強固さにのみ言及しており、発がん性（可能性）の程度には言及していない。

例えば、グループ 1:その薬剤はヒトに対して発がん性がある。このカテゴリーは、ヒトにおける発がん性の十分なエビデンスがある場合に使用される。例外的に、ヒトにおける発がん性のエビデンスが十分ではないが、実験動物における発がん性の十分なエビデンスがあり、ばく露されたヒトにおいて、その薬剤が発がん性の関連メカニズムを介して作用するという強いエビデンスがある場合には、薬剤はこのカテゴリーに分類される。国際的に広く受け入れられているが、過去には特定の薬剤の分類に対する批判があり、より最近の批判は、そのような評価のために IARC が採用した一般的なアプローチに向けられており、反論の発表を動機づける可能性がある（Pearce ら、2015 年）。

C.2. リスク評価における疫学的研究の統合に関する US-EPA の経験

米国環境保護庁の農薬プログラム（OPP）は、農薬製品の登録と規制を担当する米国の政府機関である[25]。この活動の一環として、また農薬の使用が許可される前に、OPP は農薬がヒト健康と環境に及ぼす影響を評価している。

EPA は、連邦殺虫剤・殺菌剤・殺鼠剤法（FIFRA）及び連邦食品・医薬品・化粧品法（FFDCA）を通じて、農薬製品のリスクを特性評価するための広範なハザード及びばく露情報を入手している。農薬の毒性影響に関する情報は、一般的に、農薬登録者が実施し、EPA に提出する実験動物を用いた研究から得られている。

これまでは、農薬へのばく露に関連する可能性のある潜在的なリスクを EPA が評価するための情報として、農薬に関する十分に計画された疫学研究から得られた情報は一般的には得られていなかった。農薬へのばく露と健康影響と

[25] 農薬科学及び農薬リスクの評価に関する一般的な情報については、https://www.epa.gov/pesticide-science-and-assessing-pesticide-risks を参照のこと。

の間に考えられる関連性を調査する疫学研究が文献に掲載されることが増えてきたため、EPA はこの情報源をさらに重視している。これは、20 年以上にわたって 9 万人近くの個人を追跡した大規模でよく実施された前向きなコホート研究である農業健康調査（Agricultural Health Study[26]（AHS）や、小児環境保健・疾病予防研究センターから得られた豊富な研究に特に当てはまる。EPA[27]は、このような疫学的情報を最も科学的に強固で明白性の高い方法で利用することを目標に、ヒトの健康リスク評価においてこれらの疫学的研究をより多く利用することを意図している。

C.2.1. OPP 疫学的フレームワーク文書

このプロセスの初期段階として、EPA-OPP は、2010 年に「健康リスク評価にヒトの疫学的データ及びインシデントデータを組み込むためのフレームワーク（Framework for incorporated human Epidemiologic and Incident Data in Health Risk Assessment）」（US-EPA、2010 年 a）という疫学的枠組み文書案を作成した。2010 年のフレームワーク草案は、2010 年 2 月に FIFRA 科学諮問委員会（SAP）によって好意的にレビューされた（US-EPA、2010 年 b）。この文書は最近、2016 年に「Office of Pesticide Programs' Framework Document for Incorporating Human Epidemiology and Incident Data in Risk Assessments for Pesticides」（US-EPA、2016 年）に更新された。改訂及び更新された 2016 年のフレームワーク文書は、疫学研究（ヒト事例データベース、バイオモニタリング研究に加えて）でみられるようなヒトの情報が、実験的な毒性学的情報とともに、実際の化学物質ばく露によって引き起こされる影響についての予測を提供することで、この新しいアプローチにおいて重要な役割を果たすことが提案されている。さらに、疫学的／分子疫学的データは、追加解析の指針となり、潜在的に影響を受けやすい集団や新たな健康影響を特定し、既存の毒物学的観測を補完する可能性がある。2016 年版フレームワークの概念は、専門家の査読を経た強固な原則とツールに基づいており、疫学データのレビューと評価のための多くの既存のガイダンス文書とフレームワーク（表 C.1）に依存している。また、問題設定の重要性と生物学的組織の異なるレベルでの情報統合の必要性を強調した世界保健機関／国際化学安全計画（MOA）／ヒト関連性フレームワークの更新とも一致している（Meek ら、2014 年）。さらに、このフレームワークは、2009 年の報告書「Science and Decisions（NRC、2009 年）」の中で、複雑な科学的分析の最初に問題の定式化を使用することの重要性を説明しているという点で、全米科学アカデミーの全米研究評議会（NAS／NRC）の勧告と一致している。問題の定式化の段階は、解析の目標と可能なリスク管理戦略を特定するためのリスク管理者との計画的な対話から始まると想定されている。この最初の対話は、科学的解析のための規制の背景を提供し、そのような解析の範囲を明確にするのに役立つ。問題設定の段階では、農薬の使用／使用、懸念される毒性学的影響、ばく露経路、持続時間、データや科学的情報の主要なギャップに関する利用可能な情報を考慮することも含まれている。

[26] https://aghealth.nih.gov/を参照

[27] https://www.epa.gov/research-grants/niehsepa-childrens-environmental-health-and-disease-prevention-research-centers を参照してください。

表 C.1：OPP が使用している主なガイダンス文書とフレームワーク（US-EPA、2016 年より）

	1983	連邦政府におけるリスクアセスメントプロセスの管理
NAS	1994	科学と判断
	2007	21 世紀の毒性試験
	2009	科学と政策決定。リスク評価の推進
WHO/IPCS	2001–2007	行動様式／ヒトとの関連性のフレームワーク
	2005	化学物質調整係数（CSAF）
	2014	行動様式／種のコンコーダンス解析に関する WHO/IPCS フレームワークの進化と応用における新展開
EPA	1991–2005	リスクアセスメントフォーラムリスクアセスメントのためのガイダンス（例：発がん性、生殖、発生、神経毒性、生態学的、ばく露評価のためのガイドライン、ベンチマーク用量モデリングのためのガイダンス、参照用量と参照濃度プロセスのレビュー
		http://www.epa.gov/risk_assessment/guidance.htm
	2000	リスク特性評価に関する科学政策ハンドブック http://nepis.epa.gov/Exe/ZyPURL.cgi?Dockey=40000006.txt
	2006	生理的薬物動態（PBPK）モデルのリスク評価への応用のためのアプローチとその裏付けとなるデータ
	2014	政策決定に役立つヒトの健康リスク評価の枠組み
	2014	種間・種族内の外挿のためのデータ由来の外挿係数を開発するための定量データ適用の手引き
	2001	総合的なリスク評価
		https://www.epa.gov/sites/production/files/2015-07/documents/aggregate.pdf
OPP	2001 and 2002	累積リスク評価 http://www.epa.gov/ncer/cra/
OECD	2013	経済協力開発機構（Organisation for Economic Co-operation and Development）有害性転帰経路の開発と評価に関するガイダンス文書

EPA フレームワーク文書は、このような疫学的研究及び科学的情報を農薬化学物質のリスク評価にどのように組み込むことができるかを評価する際に、また、有害性転帰経路（または MOA）の理解との関連で複数の科学的エビデンスを評価するための基盤を提供する際に、EPA が考慮する科学的考察を記述している。このフレームワークは、疫学、毒性学、リスク評価の標準的な手法に依存し、それを支持しているが、新しい情報源や追加の情報源からの情報を取り入れることも可能である。この機関のフレームワークの重要な構成要素の一つは、実験研究と観察研究の両方で観察された因果関係の特性を知るために、異なる情報源からの情報を整理して統合するためのツールとして、MOA フレームワーク／有害性転帰経路の概念を使用することである。MOA（Boobis ら、2008 年；Simon ら、2014 年；Meek ら、2014 年）と有害性転帰経路（Ankley ら、2010 年）は、フレームワーク文書で議論されている統合解析において重要な概念を提供する。MOA と有害性転帰経路の両方とも、化合物へのばく露によって引き起こされる毒性影響は、ヒトの毒性影響をもたらす一連の因果関係のある生物学的重要事象によって記述できるという前提に基づいており、その目的は、環境物質へのばく露がこれらのパスウェイをどのようにかく乱させ、それによって毒性影響につながる後続の重要事象のカスケードを引き起こすかを決定することである。

フレームワークの多くの概念は、全米科学アカデミー（National Academies, Science and Decisions）の 2 つの報告書：「リスクアセスメントの進展」（Advancing Risk Assessment：NAS 2009 年）と「21 世紀の毒性試験」（Toxicity Testing on the 21st Century: NAS 2007 年）から引用されている。これら 2 つの NRC 報告書は、毒性試験の実施方法、データの解釈方法、そして最終的に規制上の政策決定の方法を大幅に変更することを提唱している。特に、21 世紀の毒性試験に関する 2007 年の報告書では、毒性試験、リスク評価、政策決定をより良く伝えるために、現在の先毒性エンドポイントの頂点の使用に焦点を当てたものから、毒性経路を使用することへの決定的な変化を提唱している。

MOA のフレームワークは、原因経路に沿って、そして用量反応、時間的一致、生物学的妥当性、一貫性、一貫性などの要素を考慮に入れ、Bradford Hill によって記述されたものに基づいた基準を使用して、エビデンスの重み付け

に基づいて確立された一連の重要事象を特定することから始まる。特に、修正された Bradford Hill 基準(Hill、1965 年)は、MOA または有害な影響の現経路内の重要事象を立証する実験的支持を評価するために使用され、エビデンスの重み付け分析において、強度、一貫性、用量反応、時間的一致、生物学的妥当性などの概念を明示的に考慮している。この分析的アプローチを用いることで、疫学的結果は、報告された結果の一貫性、再現性、生物学的妥当性を評価し、不確実性の領域と将来の研究を特定するために、他のヒト情報との関連で評価することができる。以下の図 C.1(NRC、2007 年より引用)は、異なるタイプの情報が生物学的組織の複数のレベル(分子レベルから集団ベースのサーベイランスに至るまで)でどのように相互に関連しているかを示唆しており、遺伝子、タンパク質、低分子がヒトの細胞機能を維持する分子経路を形成するためにどのように相互作用するかという急速に発展している科学的理解に基づいている。

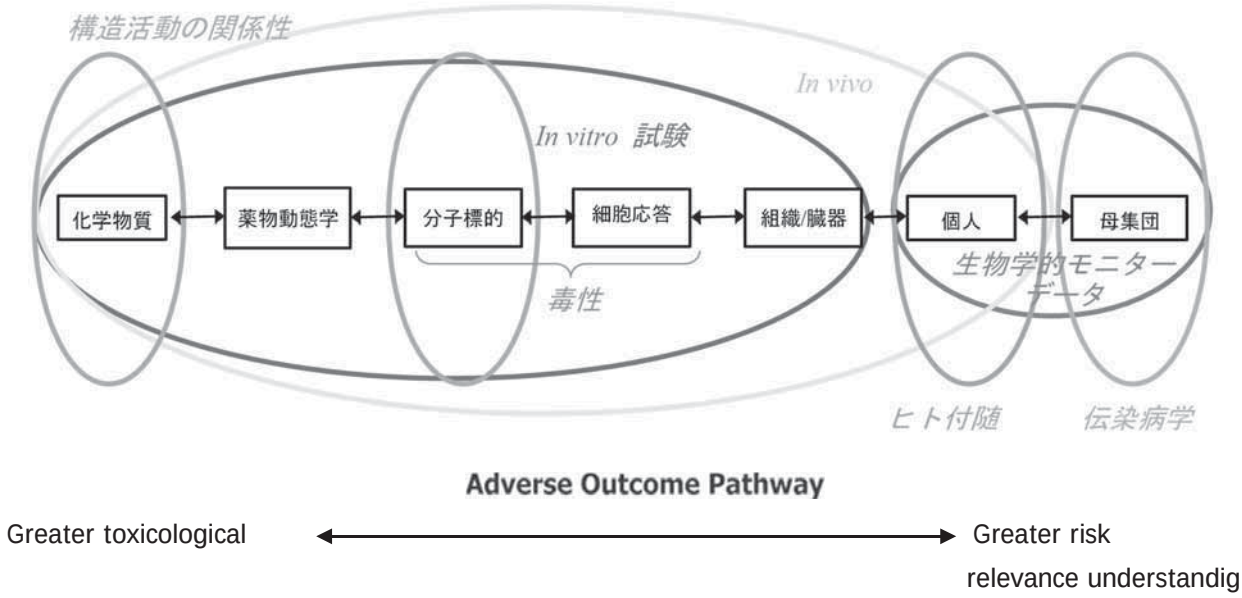


図 C.1:起源から健康影響への経路。生物組織のレベルにまたがる化学物質の影響(NRC、2007 年より引用)

C.2.2. システマティックレビュー:目的にかなった

全米アカデミーの全米研究評議会(NRC)は、EPA の IRIS プログラムのレビューにおいて、システマティックレビューを「特定の問題に焦点を当て、明確かつ事前に特定された科学的方法を用いて、類似しているが別個の研究の結果を特定、選択、評価、要約する科学的調査」と定義している[28]。近年、NRC は、規制上の政策決定に情報を提供するために化学的特異性リスク評価をサポートする科学的な文献レビューの明白性を高めるシステマティックレビュープロセスに移行することを EPA に勧めている[6]。

NRC の勧告に沿って、EPA-OPP は、政策決定を支える科学的データの収集、評価、統合のための明白性の高い方法に依拠した、目的にかなったシステマティックレビューを採用している。そのため、各システマティックレビューの複雑さと範囲はリスク評価ごとに異なる。EPA-OPP は、対象範囲/問題策定から始まり、データ収集、データ評価、データ統合及び重要なデータギャップが特定された結果の要約が行われる。

システマティックレビューでは、対象となる研究の結果をまとめるために統計的手法(メタアナリシスなど)やその他の定量的手法を使用することが多く、利用可能なエビデンスのレベルや存在する可能性のあるバイアスの程度を評価するために半定量的な採点システムを使用することができる。EPA の農薬プログラム管理局の場合、規制審査プロセスの一環として実施されるこのような Tier III(システマティックレビュー)評価には、審査中の農薬化学物質と、(最初の Tier II 評価で示唆されたように)特定の関連する健康影響が疑われる農薬化学物質の審査が含まれることになる。

米国の多くの連邦政府やその他の組織が、このようなシステマティックレビューの実施方法を評価したり、ガイダンス

[28] http://dels.nas.edu/Report/Review-Integrated-Risk/18764

文書を発行したりしており、多くのフレームワークが開発されている。これらには、EPA IRIS プログラムのアプローチ[29]、National Toxicology Programs'Office of Health Assessment and Translation（NTP/OHAT）アプローチ[30]、Cochran Collaboration のアプローチ[31]、Campbell Collaboration 及び Navigation Guide[32]が含まれ、後者については Environmental Health Perspectives 誌の一連の記事で説明されている。それぞれのアプローチは、データ収集、データ評価、データ統合、要約／更新という 4 つのステップを大まかに共有している。例えば、The Cochrane Collaboration は、The Cochrane Handbook for Systematic Reviews of Interventions for evidence-based medicine の中で、システマティックレビューの重要な主要特性の多くをリストアップしている（US-EPA、2016 年）。

- 目的が明確に示されており、研究の適格性基準があらかじめ定義されていること
- 明示的で再現性のある方法論
- 適格性基準を満たすすべての研究を特定するための系統的な検索
- 特定された研究から得られた知見の妥当性の評価
- 収録された研究の特性と知見を体系的に提示し、総合的にまとめたもの。

この付属書の以下のセクションで説明・詳述されているように、農薬リスク評価への疫学的データのレビューと統合に対する OPP のアプローチは、「作成された評価が必要な政策決定を報告するのに適しており有用であることを確認し（US-EPA、2012 年）、必要な供給源が、より詳細な研究から得られる予測または予測される情報と一致しているか、バランスが取れていることを確認する」という意味で、各段階が目的に応じて適切に実施される段階的なアプローチを採用している。Tier I 評価は、調査及び評価が AHS に由来する研究に限定されている場合の、調査実施または調査実施の更新のいずれかである。Tier II 評価では、疫学的文献の広範な検索、包括的なデータ収集、より深く、より関与したデータ評価が行われ、より広範であるが、一般的には範囲が疫学に限定されており、疫学、ヒト中毒事例、動物毒物学、有害な影響経路を横断した学際的な統合には至らない。Tier III 評価は、データ統合とより広範なデータ評価と抽出を伴う完全なシステマティックレビューであり、メタアナリシスやメタ回帰、因果関係推論／因果関係図、定量的バイアス解析や感度分析などのより高度な疫学的手法を含むことがある。

C.2.3. 現在及び将来予想される EPA 疫学レビューの実施

C.2.3.1. Tier I（scoping と問題の定式化）と Tier II（より広範な文献検索）

現在 EPA では、農薬の疫学的レビューは、上述の通り、リスク評価の進展に応じて段階的なプロセスで実施されている。この初期段階の Tier I/scoping 疫学報告書の目的は、プロセスの問題の定式化/scoping 段階で関連性の高い疫学研究が検討され、適切な場合には、プロセスの（後期の）リスク評価段階で十分に検討されるようにすることである。Tier I 段階では、EPA-OPP は、農薬問題に焦点を当てた質の高い有名なコホート研究、特に農業健康調査（Agricultural Health Study：AHS）に焦点を当てている。AHS は、農薬ばく露とがん及びその他の健康影響との関連を評価する連邦政府出資の研究であり、米国国立がん研究所（NCI）、国立環境衛生科学研究所（NIEHS）、CDC の国立労働安全衛生研究所（NIOSH）及び米国環境保護庁との共同研究を代表するものである。AHS 参加者コホートには、アイオワ州とノースカロライナ州の 89,000 人以上の免許を持つ商業及び民間の農薬散布者とその配偶者が含まれている。登録は 1993 年から 1997 年まで行われ、データ収集は現在も継続中である。AHS は、AHS コホートに関連した、またそれを利用した出版物のリストをウェブサイトに掲載している（https://aghealth.nih.gov/news/publications.html を参照）。

対象となる農薬が AHS（www.aghealth.org）の一部として調査されている場合、EPA の「scoping」解析の一環として内容摘要（または「調査資料」）が公表されるため、評価の早い段階でこれらの研究の予備的な（Tier I/scoping）レビ

[29] https://www.epa.gov/iris/advancing-systematic-review-workshop-December-2015 年を参照。
[30] http://ntp.niehs.nih.gov/pubhealth/hat/noms/index-2.html 及び NTP の「系統的レビューとエビデンス統合のための OHAT アプローチを用いた文献ベースのアセスメントを実施するためのハンドブック」https://ntp.niehs.nih.gov/ntp/ohat/pub s/handbookjan2015_508.pdf を参照。
[31] http://handbook.cochrane.org/を参照。
[32] http://ehp.niehs.nih.gov/1307175/を参照。

ューが実施される。この初期の段階では、AHS研究の基本的な疫学的結果と結論は、審査中の農薬に関連するAHS結果がある場合には、様々な AHS 研究の著者による適切な結論を簡潔に要約することを目的とした Tier I/scoping 文書に記載されている;この Tier I scoping レビューは詳細な内容、批判的な評価またはエビデンスの統合を提供するようになっていなくて AHS に関連する学術論文の要約されたハイライトに触れる可能性があるだけである。子ども環境保健・疾病予防研究センター(Children's Environmental Health and Disease Prevention Research Centres)の研究のように、AHS 以外の高品質な研究がある場合は、これらの研究も同様にこの Tier I/scoping 疫学レビューに要約されているかもしれない。繰り返しになるが、文献の批評や統合は行われていない。場合によっては、Tier I/scoping レビューでは、利用可能なエビデンスの追加的な疫学的レビューはこれ以上必要ないと結論付けられることがある。あるいは、より詳細な Tier I/update または Tier II 評価の一部として、更なるレビューが必要であると勧告する場合もある。

Tier I/update 評価は通常、Tier I/ scoping 評価の完了から1年から3年後に完了し、後述する Tier II と同様に、ヒト健康リスクアセスメント案(Draft Human Health Risk Assessment)と一緒に、またその一部として発行される。Tier I/update アセスメントでは、AHS で利用可能な文献の徹底的なレビューが行われる。Tier I/update 評価では、AHS のウェブサイトに掲載されている該当する研究を質的、叙述的な要約(報告された関連性の尺度を含む)でレビューし、要約し、評価する[33]。レビューは一般的に叙述の形式で行われ、研究の主要な側面とその結論に焦点を当て、必要に応じて EPA OPP の結論の要約及び更なる研究のための勧告に加えて EPA OPP の解説を含む。

C.2.3.2. Tier II(より広範な文献検索)

Tier II 評価は、利用可能な疫学的証拠のより完全なレビューであり、一般的には、初期の Tier I/scoping 文書が特定の懸念の可能性を示唆している場合にのみ実施される(例えば、特定で信頼できるばく露-疾病仮説が進められており、より詳細な評価の一部としてさらに評価する必要がある)。Tier II 疫学評価は、Tier I/update と同様に、一般的に Tier I 評価の完了から 1 年から 3 年後に完了し、OPP のヒトの健康リスク評価案(Draft Human Health Risk Assessment)と一緒に、またその一部として発行される;Tier II 評価はシステマティックレビューの特定の要素を組み込む定性的で叙述的なレビューだと考えられる。例えば、Tier II 評価では、AHS データベースだけでなく、PubMed、Web of Science、Google Scholar、Science Direct などのデータベースや、場合によっては標準化された明白で再現性のある照会言語を使用して、専門の図書館や情報科学の支援を得て、Tier I 評価よりも広範囲に及ぶ徹底した完全な文献検索が含まれている[34]。EPA によるエビデンス統合(一般的には定性的かつ叙述的な形式で行われるが)も Tier II 評価では行われ、疫学的文献に関する全体的な結論が出されている。さらに、Tier II 評価は、特別な仮説としてばく露-健康影響に関する更なる疫学的データや研究が将来の研究のための興味深い分野を示す可能性がある。Tier II 評価文書は一般的に、疫学的知見を、動物毒性試験や、リスク評価の一部としてある程度(別個にに)行われる MOAs/AOPs からの情報などの他のエビデンスと統合しようとはしない。特定の農薬に関連すると考えられる特定の健康影響が特定される範囲までの Tier II 評価に対して、その後のより包括的な Tier III 評価では、分野を超えた更なる調査及び統合を行うことができる(下記参照)。

C.2.3.3. Tier 3(データ統合を伴う完全なシステマティックレビュー)

Tier II 評価では、ある農薬化学物質との関連性があると仮説が立てられている疫学的文献に現れる広範な健康影響を検討するが、Tier III 評価では、より広範な(学際的な)疫学的根拠に基づく、時にはより定量的/統計学的な評

[33] https://aghealth.nih.gov/news/publications.html

[34] 疫学とバイオモニタリング/ばく露の項目の下で行われた追加検索は、NHANES Exposure Reports (http://www.cdc.gov/exposurereport/); TOXNET (http://toxnet.nlm.nih.gov/); CDC NBP Biomonitoring Summaries (http://www.cdc.gov/biomonitoring/biomonitoring_summaries.html); ICICADS (http://www.inchem.org/pages/cicad s.html); ATSDR Toxicological Profiles (http://www. atsdr.cdc.gov/toxprofiles/index.asp)、IARC モノグラフ (http://monographs. iarc.fr/ENG/Monographs/PDFs/; EFSA's Draft Assessment Report Database (http://dar.efsa.europa.eu/dar-web/provision); and Biomonitoring Equivalents (https://blog.americanchemistry.com/2014/07/biomonitoring-equivalents-a-valuable-scientific-tool- for-making-better-chemical-safety-decisions/)

価を行い、これを動物毒性学や MOA/AOP 情報とより正式に統合しようとするものである。このような Tier III 評価は、疫学的文献のシステマティックレビューの形をとり、毒性及び有害な転帰の経路の評価と併せて実施される。

AHS 由来の農薬化学物質については、Tier III 解析では、他の質の高い疫学調査の評価結果を取り入れ、より多様な情報源を再調査するために「エビデンスの重み付け」をより多く取り入れることが理想的である。これらの調査の結果は、AHS の結果との再現性と一貫性を評価するために使用される。多くのケースにおける初期の AHS の結果は、特定の健康影響を示した少数の参加者、または参加者を追跡した年数が比較的少ない参加者に基づいている。AHS コホートの高齢化に伴い、AHS からのいくつかの化学物質の 2 回目の評価の発表は、さらなる追跡調査の年数と、ばく露と健康影響の間の正負の関連を解釈するためのより強固な根拠となると期待されるより多くの症例数に基づいて行われることになるであろう。さらに、AHS では、疫学研究の結果の解釈を助けるために、かなりの量の生化学的、遺伝的マーカー及び分子データの生成が増加している。このような結果は、AHS の結果をさらに明確にしたり、ばく露と健康影響を結びつける生物学的基盤のエビデンスを提供したり、あるいは因果関係の経路の基礎となるメカニズムのエビデンスを強化する可能性のある追加の実験研究や観察研究を示唆したりする可能性がある。さらに Tier III 解析では、AHS がメンバーである国際農業コホートコンソーシアム（AgriCOH）の国際的なコホート研究からの情報と結果をまとめる努力を利用することもできる。AgriCOH は、研究間のデータを蓄積するための機会と方法を特定するために積極的に取り組んでおり、これらの他のコホートデータの利用可能性は、EPA が疫学的データを検討、評価、重み付けする際に、ばく露ー健康影響の関係の再現性と反復を評価するのに役立つはずである。

C.2.4. OPP の公表文献検索戦略及び研究の質の評価

システマティックレビューアプローチの重要な側面は、確立された適格基準を満たす多くの文献をみつけることができるように、公表されている疫学的文献を徹底的に、体系的に、再現可能に検索することである[35]。OPP は文献検索の一部として特定のデータベースを使用しており、その実施に関する特定のガイダンスがある（例えば、ヒト健康リスク評価に関する OPP の公表文献検索ガイダンス[36]）。すべての関連文献の評価、エビデンスの強固さを評価するための標準化されたアプローチの適用、明確で一貫性のある総括的表現は、一般的に重要な要素となる（NRC、2011 年）。さらに、質の高いばく露評価は、環境疫学研究や職域疫学研究において特に重要である。

上記のシステマティックレビューのアプローチの第二の重要な要素は、同定された研究から得られた結果の妥当性の評価である。一般的に言えば、疫学研究の質、研究の文書化の妥当性（研究の計画と結果）、リスク評価との関連性は、政府機関のリスク評価に使用するために公表されている文献から疫学研究を評価する際に考慮される。個々の研究の質を検討する際には、疫学研究の計画、実施、解析、解釈の様々な側面が重要である。これらには以下が含まれる（US-EPA、2016 年より）。

1）仮説を明確に明示することで、たとえその研究が本質的に仮説生成的なものであったとしても、その仮説を明確に示されていること。
2）健康影響の関連する臨界期、リスク評価対象集団の関係あるばく露範囲、試験から得られる用量／ばく露反応の傾向の入手可能性など、ばく露評価の資質の中で、適切なばく露評価が十分であること。
3）合理的に有効で信頼性の高い結果の確認（研究集団における健康影響の有無を正しく識別されていること）。
4）対象集団を代表するサンプル集団となり、系統的な偏りがない適切な組み入れ基準と除外基準。
5）観察されたリスク推定値における複数の農薬ばく露、または混合物ばく露の役割の評価または考察を含む、潜在的な交絡変数の適切な評価及び解析。

[35] 出版バイアスに関連する潜在的な問題を軽減するために、はっきりしない文献や未発表の文献を見ることを提唱する者もいる。
[36] 疫学的データに関する特別な注意事項については、https://www.epa.gov/pesticide-science-and-assessing-pesticide-risks/guidance-identifying-selecting-and-evaluating-open 及び 2012 年 8 月 28 日付けの文書「Guidance for Considering and Using Open Literature Toxicity Studies to Support Human Health Risk Assessment」の 10 ページを参照のこと。

6）参加者の選択や情報収集における誤りを含む、研究における潜在的な系統的な偏りの全体的な特性。これには、提示されたリスク推定値に対する系統的誤差の潜在的な影響を調査するための感度分析の実施を含む。

7）ばく露-健康影響評価のための適切な統計的検出力、または観察された影響に対して検出力が不足している場合の研究の統計的検出力の影響の評価及び検出力推定値の適切な考察及び／または提示。

8）研究デザインと対象となる結果の性質を考慮した適切な統計的モデル化技術の使用。

参考文献

Ankley GT, Bennett RS, Erickson RJ, Hoff DJ, Hornung MW, Johnson RD, Mount DR, Nichols JW, Russom CL, Schmieder PK, Serrrano JA, Tietge JE and Villeneuve DL, 2010. Adverse outcome pathways: a conceptual framework to support ecotoxicology research and risk assessment. Environmental Toxicology and Chemistry, 29, 730–741.

Boobis AR, Cohen SM, Dellarco V, McGregor D, Meek ME, Vickers C, Willcocks D and Farland W, 2006. IPCS framework for analyzing the relevance of a cancer mode of action for humans. Critical Reviews in Toxicology, 36, 781–792.

Boobis AR, Doe JE, Heinrich-Hirsch B, Meek ME, Munn S, Ruchirawat M, Schlatter J, Seed J and Vickers C, 2008. IPCS framework for analyzing the relevance of a noncancer mode f action for humans. Critical Reviews in Toxicology, 38, 87–96.

Hill AB, 1965. The environment and disease: association or causation? Proceedings of the Royal Society of Medicine, 58, 295–300.

Meek, ME, Boobis A, Cote I, Dellarco V, Fotakis G, Munn S, Seed J and Vickers C, 2014. New developments in the evolution and application of the WHO/IPCS framework on mode of action/species concordance analysis. Journal of Applied Toxicology, 34, 595–606.

Meek, ME, Palermo CM, Bachman AN, North CM and Lewis RJ, 2014. Mode of action human relevance (species concordance) framework: evolution of the Bradford Hill considerations and comparative analysis of weight of evidence. Journal of Applied Toxicology, 34, 1–18.

NAS (National Academy of Sciences), 2007. Toxicity Testing on the 21st Century: A Vision and a Strategy. Board on Environmental Studies and Toxicology. Available online: https://www.nap.edu/catalog/11970/toxicity-testing- in-the-21st-century-a-vision-and-a

NAS (National Academy of Sciences), 2009. Science and decisions: advancing Risk Assessment. Board on Environmental Studies and Toxicology. Available online: http://dels.nas.edu/Report/Science-Decisions-Advanc ing-Risk-Assessment/12209

NAS (National Academy of Sciences), 2011. Review of the Environmental Protection Agency's Draft IRIS Assessment of Formaldehyde. Board on Environmental Studies and Toxicology. Available online: https://www.nap.edu/download/13142

Simon TW, Simons SS, Preston RJ, Boobis AR, Cohen SM, Doerrer NG, Crisp PF, McMullin TS, McQueen CA and Rowlands JC, 2014. The use of mode of action information in risk assessment: Quantitative key events/dose response framework for modelling the dose-response for key events. Critical Reviews in Toxicology, 44 (Suppl 3), 17–43.

US-EPA (Environmental Protection Agency), 2010a. Draft Framework for Incorporating Human Epidemiologic and Incident Data in Health Risk Assessment. Presented to FIFRA Scientific Advisory Panel on February 2-4 2010a. January 7. Available online: https://www.regulations.gov/document?D=EPA-HQ-OPP-2009-0851-0004

US-EPA (Environmental Protection Agency), 2010b. Transmittal of Meeting Minutes of the FIFRA Scientific Advisory Panel Meeting on the Draft Framework and Case Studies on Atrazine, Human Incidents, and the Agricultural Health Study: Incorporation of Epidemiology and Human Incident Data into Human Health Risk Assessment. MEMORANDUM dated 22 April, 2010b. SAP Minutes No. 2010-03. Available online: https://www.re gulations.gov/document?D=EPA-HQ-OPP-2009-0851-0059

US-EPA (Environmental Protection Agency), 2012. Office of the Science Advisor. Risk Assessment Forum. Draft Framework for Human Health Risk Assessment to Inform Decision Making. July 12, 2012.

US-EPA Environmental Protection Agency), 2016. Office of Pesticide Programs' Framework for Incorporating Human Epidemiologic and Incident ata in Risk Assessments for Pesticides. December 28, 2016. Available online: https://www3.epa.gov/pesticides/EPA-HQ-OPP-2008-0316-DRAFT-0075.pdf

付属書 D－影響量の拡大／膨張

この文書の本文で説明されているように、研究の検出力が低い場合には、バイアスの潜在的な原因が生じる可能性がある。このあまり知られていないタイプのバイアスは、「影響量の規模」として知られています。一般的に、小規模で検出力の低い研究では、研究の検出力が意味のある影響量を確実に検出するには不十分であるため、偽陰性が生じる可能性があることは広く知られているが、推定された影響が重要、関連性がある、または「発見された」と判断されるために、統計的閾値（例えば、統計的有意性に使用される一般的な $p<0.05$ の閾値）を通過する必要がある場合、これらの研究が影響量の膨張をもたらす可能性があることはあまり知られていない。この影響は、影響量の拡大、「勝者の呪い」、真実の膨張、または影響量の膨張として様々に知られているが、これは、「発見された」関連性（すなわち、意味があると判断されるために統計的有意性の所定の閾値を通過した関連性）が、その発見を行うために最適ではない検出力を持つ研究から得られる現象であり、その結果、意図的かつ体系的に膨張した影響量が生じることになる。

このような真偽の不明確化は、検出力の低い研究で統計的有意差を達成する研究では、帰無値から離れる（系統的な）バイアスとして現れる（Reinhart、2015 年）。これは、検出力の低い（したがって一般的には小さい）研究は、結果が大きく変動する可能性が高く、大規模な研究よりも個人の間のランダムな変動の影響を受けやすいからである。より具体的には、どのような研究でも観察されるかもしれない影響量の拡大の変化の程度は、研究の結果がどの程度広く変化すると予想されるかに部分的に依存しており、これは研究の検出力に依存する；検出力の低い研究は影響量の拡大の程度を大きくする傾向があり、その結果検出力の高い研究よりも統計的有意性（または他の閾値基準をパスする）を見出す。

この「影響量の拡大」の概念とその理由の例として、可変のサンプルサイズで何千回も試験を実施した場合を想像すると便利である。この場合、観察された影響量の広い分布がある。これらの推定影響量の観察された中央値は、真の影響量に近いと予想されるが、小規模な試験では、大規模な試験に比べて、観察された影響量のばらつきが必然的に大きくなる。しかし、低検出力の研究では、観察された影響のうち、統計的に有意な（高い）閾値を通過するのはごく一部であり、これらの影響は最大の影響量を持つものだけである。このように、一般的に小規模で、ランダム変動が大きい低検出力の研究では、与えられた統計的閾値を通過した結果、実際に有意性起因の関連を発見した場合、その影響の大きさを過大評価する可能性が高くなる。これが意味するのは、低検出力で統計的に有意な研究の結果は、膨張効果となるように偏っているということである。Gelman 及び Carlin（2014 年）が要約しているように、「研究者が小さな影響を研究するために小さな[検出力不足][37]サンプルとノイズの多い測定を使用した場合、有意な結果はしばしば驚くほど間違った方向に行き、影響を大幅に過大評価する可能性が高い」のである。一般的に、バックグラウンド（または対照または無処置）率が低い、対象となる影響量が小さい、研究中のサンプルサイズが小さいと、研究の検出力が低下し、その結果、（あらゆる）膨張した影響量の傾向と規模が大きくなることが示されている。

影響量の膨張現象は、発見科学全般に適用される一般的な原則であり、疫学の特殊な現象や弊害ではないことに注意することが重要である（Ioannidis、2005 年；Lehrer、2010 年；Button、2013 年；Button ら、2013 年；Gelman 及び Carlin、2014 年；Reinhart、2015 年）。これは、薬理学の研究、遺伝子研究、心理学の研究、そして最もよく引用される医学文献の多くでよくみられる。ほとんどの疫学研究のように、研究者がサンプルサイズを増加させる能力が限られている場合、影響効果量の拡大は、研究や研究デザインの機能や欠陥ではなく、むしろ、その研究の結果がユーザーコミュニティによってどのように解釈されるかという機能である。したがって、疫学研究における選択や情報バイアスのような他のバイアスの可能性とは異なり、バイアスは研究やその計画に内在するものではなく、むしろその研究がどのように解釈されるかに特徴がある。

統計的に有意な結果をもたらす研究について、影響量の規模の潜在的な程度を決定（定量化）するために、査読者は様々な検出力の計算を実行しなければならない。より具体的には、化学物質ばく露と疾患との間の関連が統計的に有意であることが判明した場合、検出力解析は、統計的に有意な影響量の推定値（例えば、オッズ比、相対リスクまたは率比）がどの程度人工的に膨張しているかを決定するために行われる。

[37] [斜体を付けた]

必要な検出力計算を行うために、査読者は以下の 4 つの値を知っているか、または得なければならない。

1）非ばく露群の被験者数。

2）ばく露群の被験者数。

3）非ばく露群の対象疾患を持つ個人の数（または症例数）及び

4）2 つのグループ（例えば、ばく露群 vs.非ばく露群）の比較において、所定の（事前決定の）量の差を検出するための対象となる目標値。

最初の 3 つの値は文献に記載されているか、文献から入手しなければならないが、対象となる目標値（一般的に疫学研究では OR または RR）はリスク管理者によって選択される（最終的には政策決定である）[38]。本付属書では、シミュレーションを用いて、この影響量の膨張現象を定量的に検討する。本付属書では、2 つの公表された研究例と数百件の試験のシミュレーションを用いて、検出力が低いために偏った影響量（オッズ比、率比、相対リスクなど）を生み出すのに影響量の拡大がどの程度の役割を果たしているかを評価している。

最初の例では、ダイアジノンばく露と肺がんを調査した Agricultural Health Study の前向きコホート出版物からのデータを使用し、計算された RR の影響量の偏りの問題を説明している。2 番目の例は、マラチオンばく露と NHL を研究した症例対照研究からの ever-never データを使用して、推定 OR の観点から影響量の拡大の概念を説明する。

影響量の拡大の説明と相対リスクを説明する例（Jones ら、2015 年）

ダイアジノンにばく露されていないものと、最も高い三分位（T）でばく露されているものとの間の比較に関連した検出力は、肺がんに対する Jones ら（2015 年）の AHS 研究発表「Incidence of solid tumours among pesticide applicators exposed to the organophosphate insecticide diazinon in the Agricultural Health Study - an updated analysis」で提供された情報から計算することができる。各ばく露量での被験者数は文献中に記載されていた（非ばく露群。N=17710 及び T（ertile）1、T2 及び T3 は、ばく露分布に基づいて分類された；具体的には、各三分位の N=（2,350 + 2,770）/3=1,710 文献の表 1 から。（a）2,350 の値は最も低いばく露量を表し、（b）2,770 の値は、ばく露された被験者を二分法で分類したときの、2 つの最も高いばく露量を表している。（i）参照非ばく露群の被験者数=17,710；（ii）ばく露群（三分位）のそれぞれの被験者数=1710；（iii）参照非ばく露群の罹患者数（肺がん）=199 人（引用文献の表 3 より）とすると、真の率比=1.2、1.5、または 2.0 と仮定すると、文献に示された T1 対非ばく露群、T2 対非ばく露群、T3 対非ばく露群の比較の検出力を計算することができる。

ここで、我々は、199/17710（=0.011237）の推定バックグラウンド率及び感度分析の形態として、このバックグラウンド率の 1/2（または 0.005617）及びダイアジノンにばく露された個人の各三分位の被験者の間で 1.2、1.5、2.0 及び 3.0 の（可能な規制関係の）相対率を検出するためのこの率の 2 倍（0.022473）に関連する検出力を評価することに注目している。この解析は Stata 統計ソフトウェアを使用して行われ、1.2、1.5、2.0 及び 3.0 の真の率比について、199 人の罹患者/17,710 人のバックグラウンド率に対して 1/2x-、1x-（太字／網掛けで以下に示す）及び

[38] この目標値は対象となる効果量であり、相対リスク（コホート研究の場合）またはオッズ率（症例対照研究の場合）のいずれかで表されることが多い。すなわち、目標値は一般的に、リスク管理者が一定の確信度で検出したいと考えている一定の大きさの OR または RR である。OR または RR が高ければ高いほど、ばく露と健康影響との間の推定関連性の規模が大きくなる。何が「弱い」関連性と「強い」関連性を構成するかについての厳密なガイドラインはないが、約 1 以下（時には 1.2 以下）の値は「帰無」または「本質的に帰無」と考えられる（これは、いくつかの背景（例えば、ワクチン接種の有効性など）では考慮することが適切であるかもしれない保護効果の可能性を無視している）。2 または 3 未満の値は、しばしば「弱い」と考えられている。2（または 3）より大きく約 5 までの値は「中等度」と考えられ、5 より大きい値は「大」であると考えられている。Monson（1990）は、関連性の強さの目安として、1.0～1.2 を「なし」、1.2～1.5 を「弱」、1.5～3.0 を「中程度」、3.0～10.0 を「強」としている。他の著者は疫学においては Cohen の基準を用いて、1.5 を「小」、5 を「大」、3.5 を「中」と表現している（Cohen and Chen、2010）。また、1.5 を「小」、2.5 を「中」または「中程度」、4 を「大」または「強」、10 を「非常に大きい」または「非常に強い」と表現する人もいる（Rosenthal、1996 年）。Taube（1995 年）は、弱い関連性を検出する上での環境疫学の限界について議論している（Wynder（1997 年）の反論を示す招待解説も参照のこと）。これらの境界線はどれも「難しい」ものではなく、どこに線が引かれ、どのように考えられ、どのように解釈されるかについては、正当な意見の相違があり得ることを認識すべきである。それにもかかわらず、これらの境界線は背景に大きく依存するものであり、上記のような境界線は、いかなる意味でも、公式なものでも、決定的なものでもあると考えるべきではない。

2x-（観察された）を表形式及びグラフ形式の両方で以下に示す[39]。

片側 2 標本の比率検定の検出力解析結果（a = 0.05）(a)

$N_{control}$	$N_{exposed}$	Proportion control(b)	Proportion exposed	Relative risk	Power
17,710	1,710	0.00562	0.00674	1.2	0.1634
17,710	1,710	0.00562	0.00843	1.5	0.4353
17,710	1,710	0.00562	0.01124	2.0	0.8182
17,710	1,710	0.00562	0.01685	3.0	0.9935
17,710	**1,710**	**0.01124**	**0.01348**	**1.2**	**0.2259**
17,710	**1,710**	**0.01124**	**0.01685**	**1.5**	**0.6379**
17,710	**1,710**	**0.01124**	**0.02247**	**2.0**	**0.9652**
17,710	**1,710**	**0.01124**	**0.03371**	**3.0**	**1**
17,710	1,710	0.02247	0.02697	1.2	0.3353
17,710	1,710	0.02247	0.03371	1.5	0.8632
17,710	1,710	0.02247	0.04495	2.0	0.9991
17,710	1,710	0.02247	0.06742	3.0	1

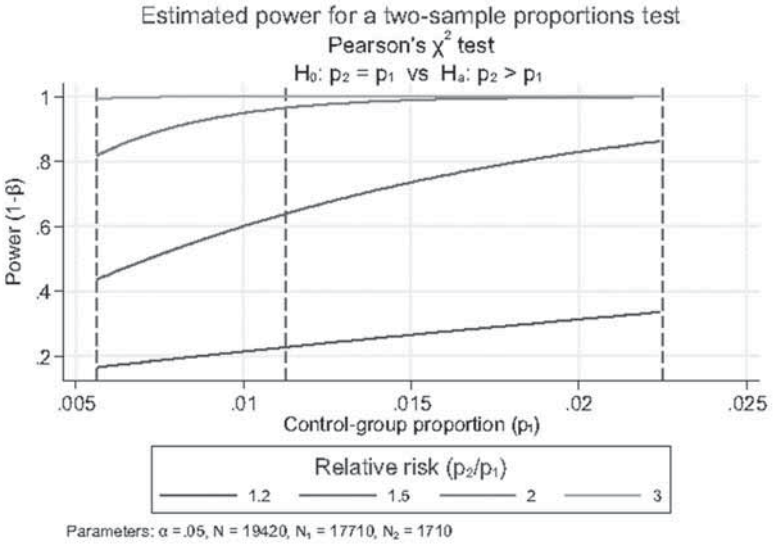
上記の検出力計算結果を生成するために使用される Stata コード：検出力 2 比例('= 0.5 * 199/17710'= 199/17710'= 2 * 199/17710), test(chi2) RR (1.2 1.5 2.0 3.0) n1(17710) n2 (1710)片側表(N1:"N コントロール" N2:"N エクスポージャー" p1:"割合コントロール" p2:"割合エクスポージャー" RR:"相対リスク" power:"検出力')。

(a)：片側検定 a＝0.05 Ho：p2＝p1 vs Ha：p2＞p1；Ncontrols＝17,710、Nexposed＝1,710；反復回数＝1,000 回（データセット）。

(b)：Jones ら（2015）の 199/17710 の肺がんの観察されたバックグラウンド率を 1/2 倍-、1 倍-、2 倍-を表す。

上表の強調表示／太字の領域は、引用研究における肺がんのこの 1x 観測されたバックグラウンド率に関連した検出力を表している。

これらの値は以下のようにグラフ化することができる[40]。



[39] 1.2、1.5、2.0、3.0 の RR は、リスク管理者や意思決定者が関心を持ちそうな一連の相対リスクに関連する検出力検出力を示すために、やや恣意的に選択されたものである。RR または OR=2.0 と 3.0 の値は、弱い影響量の大きさと強い影響量の間の境界線であると考えられている。RR 値 1.2 は、一部の人が「帰無に近い、または本質的に帰無」と考えるものであり、RR 値 1.5 はこれらの間の中間値である。疫学的証拠がばく露と健康影響との間の関係を示唆しているかどうかを判断する際に、リスク管理者は、許容可能な統計力（一般的には 80-90%と考えられている）を持つ強固な研究から得られた「本質的には帰無」RR1.2 を、関連性を見いだせなかったことを示す十分なエビデンスであり、事実上、ばく露と健康影響との間に観察可能な関連性がないという結論を支持するエビデンスを提供していると判断されることがあるかもしれない。

[40] 上のグラフを生成するための Stata コード：乗 2 比例(' = 0.5 * 199/17710'(0.0001) '= 2 * 199/17710'), test(chi2) rrisk(1.2 1.5 2.0 3.0) n1(17710) 0) n1(17710) n2(1710)グラフ(recast(line) xline('= 0.5 * 199/17710'' '= 199/17710'' '= 2 * 199/17710', lpattern(dash)) legend(rows(1)size(small)) ylabel(0.2(0.2)1.0))片側グラフである。

1.2-、1.5-、2.0-及び 3.0 の真の RR における制御群比率の関数としての検出力を評価する(片側)2 標本比率検定の推定検出力を示すグラフ。検定では、1.2-1.5-2.0、2.0-3.0 の真の RR でのコントロールグループ比率の関数としての検出力を評価している。

上の表とグラフからわかるように、本研究では、バックグラウンド率(対照比率、199 人の罹患者/17,710 人=0.011237)の 1 倍で約 23%の検出力で RR1.2 を検出することができた。約 64%の検出力が 1.5 の RR を検出する。真のバックグラウンド率が実際には観測されたバックグラウンド率の 2 倍(2 9 0.011237=0.022473)であれば、RR1.5 を検出できる検出力は約 86%、RR2.0 を検出できる検出力は実質的に 100%となる[41]。

上記を考えると、真の相対リスクを 1.2、1.5、2.0、3.0 とした場合に、どの程度の影響量の拡大(別名、影響量の膨張)があるかをシミュレーションするために SAS が使用された。下の表は、ダイアジノンと肺がんの検出力解析を示しており、シミュレーション結果から影響量の拡大の程度を示している。下の表に示された解析は、Ioannidis(2008 年)によって行われたものと類似していて、彼の表 2 に示された、影響量の拡大の概念を説明するための正式な統計的有意性の閾値を通過した仮説的な結果のセットである。

1.2, 1.5, 2.0 及び 3.0 の真のオッズ比が与えられた場合の影響量の拡大の説明を示す SAS シミュレーション結果 (a)

真値		解析したデータセット数 (N)	検出力(b)	観測された有意な RR の分布			
対照における罹患者の割合	RR			N	10th percentile	Median (% inflation)	90th percentile
0.005617 (1/2 9 background)	1.2	1,000	0.16	157	1.6	1.7 (42)	2.0
	1.5	1,000	0.4	401	1.6	1.8 (20)	2.3
	2	1,000	0.82	823	1.7	2.1 (5)	2.8
	3	1,000	1	997	2.3	3.0 (0)	3.9
0.011237 (1 9 background)	**1.2**	**1,000**	**0.22**	**224**	**1.4**	**1.6 (33)**	**1.8**
	1.5	1,000	0.63	627	1.4	1.6 (7)	2.0
	2	1,000	0.98	977	1.6	2.0 (0)	2.5
	3	1,000	1	1,000	2.5	3.0 (0)	3.6
0.022473 (2 9 background)	1.2	1,000	0.33	331	1.3	1.4 (17)	1.6
	1.5	1,000	0.87	871	1.3	1.5 (0)	1.8
	2	1,000	1	1,000	1.7	2.0 (0)	2.3
	3	1,000	1	1,000	2.6	3.0 (0)	3.4

ポアソン回帰モデルが，グループ間の（相対リスク）率を比較するために使用された．EXACT 検定は、一般化ヘシアン行列が正のデフィニートでない場合（グループの 1 つでゼロのケースがあるため)、いくつかのデータセットの解析に使用された。

(a)：片側検定、a=0.05、N コントロール=17,710、N ダイアジノンばく露者=1,710、反復回数=1,000 回（データセット)。

(b)：このシミュレーションから得られた検出力は、SAS (PROC POWER)や Stata (power two-proportion)のような統計ソフトの組み込まれた手順から計算された検出力に近いかもしれないが、正確には一致しないかもしれない。これは、シミュレーションされたデータセットの数が十分でないためかもしれない。しかし、1,000 回の反復は、検出力を適切に推定し、統計的に有意な結果（ここでは、≤0.05）を与えられた影響量の拡大の程度を説明するのに十分である。

p<0.05 で統計的に有意な結果が得られた場合、統計的に有意な結果の中央値での影響量の拡大の変化率は、ダイアジノンにばく露されていない個人の肺がんの割合(すなわち、非ばく露グループの罹患者の割合)と真の相対リスク(1.2 から 3.0 までの範囲)の両方に応じて 0%から 42%まで変化することに注意する。例えば、ばく露対非ばく露の三分位の真の RR が 1.2 であった場合、非ばく露グループの肺がんの割合は 0.011237(上記の表の太字の行)で、観察された統計的に有意な RR の半分は 1.6 の中央値を超えており、半分は 1.6 以下になるだろう;これは、シミュレーションで使用される 1.2 の真の RR の上に 33%の膨張があることを表している。

Jones ら(2015 年)の研究(0.011237)でみつかったバックグラウンド率については、統計的に有意であることが判明

[41] 別の言い方をすると、真の(しかし未知の)バックグラウンド率が実際に観測されたバックグラウンド率の 2 倍であった場合、統計的に有意な関係がみつからなかった場合、我々は合理的に(86%の信憑性で)真の OR が 1.5 を超えていないと結論付けることができる。

した真の RR の 1.2 は、前述の中央値 1.6 の代わりに 1.4（10%のパーセンタイルで）から 1.8（90%のパーセンタイルで）に変化することが観察される研究が繰り返されるだろう。真の RR が 2 または 3 のときには、検出力は 80%以上（上記の表に見られるように）であり、観測された RR の中央値は真の RR に近く、観測された RR の範囲は狭い。真の RR が 3 になると、影響量の膨張がなくなり、シミュレーションの中央値が真の RR に回帰するように検出力が増加する。

エバー／ネバー解析における影響量の拡大とオッズ比を示す例（Waddell ら、2001 年）

ばく露群と非ばく露群の比較は、三分位や四分位などの他の分類やグループ分けに基づいた比較とは対照的に、「これまでありと、決してない」の比較で示されることがある。このようなばく露カテゴリーベースの解析は、ばく露カテゴリーを小さな（より均質な）ばく露分類やグループに分けるには十分な数の症例数がないため、あるいはばく露の測定値が利用できないか、あるいは信頼性が低いため（症例対照研究のような）に行われるかもしれない。これらの状況では、（i）非ばく露群の被験者の総数、（ii）ばく露群の被験者の数、（iii）ばく露群と非ばく露群の間の比較の検出力を計算するために、非ばく露群の罹患者の数、（iv）与えられた、または事前に選択されたオッズ比が同様に必要となる。

エバー／ネバー分類を用いた症例対照研究において、検出力と影響量の拡大の解析がどのように行われるかを説明するために、マラチオンと NHL の関連性を調査した研究（Waddell ら、2001）を選択した。ここでは、（i）基準非ばく露群の被験者数=1,018 人（表 1 より：非農家=243 人の罹患者+775 人の非罹患者）、（ii）ばく露群の被験者数＝238 人（表 4 より：マラチオンばく露者=91 人の非ばく露者+147 人の非ばく露コントロール）、（iii）参照非ばく露群の罹患者数=243 人（表 1 より：非農家または非ばく露群の 243 人の罹患者）とすると、真の率比=1.2、1.5、または 2.0 と仮定すると、「ばく露ありと、ばく露なし」の比較の検出力が計算できた。

肺がんとダイアジノンについて上述したように、非農家（非ばく露者）では、以下の表に示すように、研究で推定された NHL の割合が 0.2387 で 1.2 の OR を検出する検出力を 30.5%と推定した。

片側 2 標本の比率検定の検出力解析の結果（a = 0.05）[(a)]

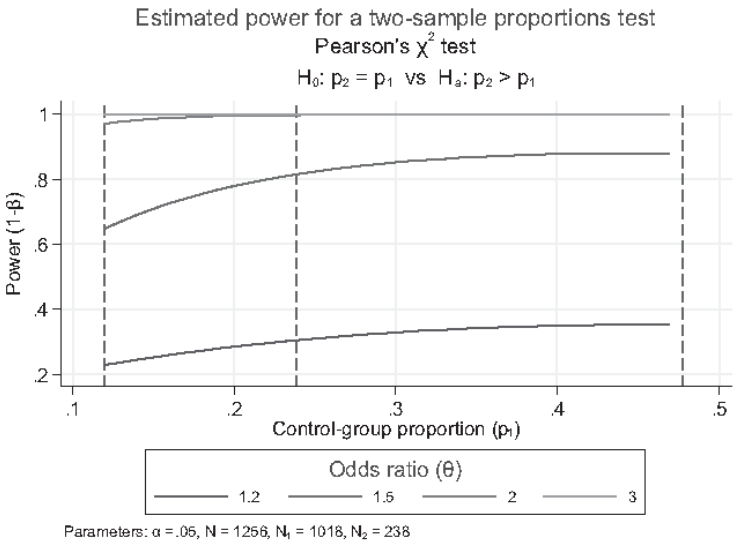
$N_{control}$	$N_{exposed}$	Proportion control[(b)]	Proportion exposed	Odds Ratio	Power
1,018	238	0.1194	0.1399	1.2	0.2279
1,018	238	0.1194	0.1689	1.5	0.647
1,018	238	0.1194	0.2133	2.0	0.9693
1,018	238	0.1194	0.2891	3.0	1
1,018	**238**	**0.2387**	**0.2734**	**1.2**	**0.3047**
1,018	**238**	**0.2387**	**0.3199**	**1.5**	**0.8149**
1,018	**238**	**0.2387**	**0.3854**	**2.0**	**0.9971**
1,018	**238**	**0.2387**	**0.4847**	**3.0**	**1**
1,018	238	0.4774	0.523	1.2	0.3522
1,018	238	0.4774	0.5781	1.5	0.8779
1,018	238	0.4774	0.6463	2.0	0.9992
1,018	238	0.4774	0.7327	3.0	1

上記の結果を生成するために使用される Stata コード：2 乗比例('= 0.5 * 243/1018' '= 243/1018' '= 2 * 243/1018'), test(chi2) OR (1.2 1.5 2.0 3.0) n1(1,018) n2(238) 片側
table(N1:"N コントロール" N2:"N ばく露" p1:"割合コントロール" p2"割合ばく露" OR:"オッズ比" power:"検出力")
(a)：片側検定 a = 0.05 Ho: p2 = p1 vs Ha: p2 > p1; Ncontrols = 1,018、Nexposed = 238、反復回数 = 1,000 回（データセット）。
(b)：Waddell ら（2001）の 243/1018 の肺がんの観察されたバックグラウンド率を 1/2 倍、1 倍、2 倍を表す。上表のハイライトされ太字の領域は、引用された研究におけるこの 1 倍の NHL の観測されたバックグラウンド率に関連した検出力を表している。

このようなマラチオンと NHL の検出力関係は、上記の AHS 前向きコホート研究（ダイアジノンと肺がん）と同様に以下のようにグラフ化されている[42]。グラフの中央縦点線は非農家／非ばく露群の NHL 比が 0.2387 での検出力を示し、

[42] グラフを生成するための Stata コード：2 乗比例 ('= 0.5 * 243/1018'(0.01) '= 2 * 243/1018'), test(chi2) OR (1.2 1.5 2.0 3.0) n1(1018) 0) n1(1018) n2(238)graph(再キャスト(線) x-line('= 0.5 * 243/1018'' '= 243/1018'' '= 2 * 243/1018', lpattern(dash)) legend(rows(1)size(small)) y-label(0.2(0.2)1.0))片側。

左端と右端の縦点線は非農家／非ばく露群の NHL 比が 1/2 と 2 倍での感度分析の枠を示している。



1.2-、1.5-、2.0-及び 3.0 の真の RR におけるコントロール群比率の関数としての検出力を評価する（片側）2 標本比率検定の推定検出力を示すグラフ。赤い破線の縦線は、観測された比率の 1/2 倍、観測された比率の 1 倍及び観測された比率の 2 倍におけるコントロールグループの比率を表し、これらのバックグラウンドレートの仮定に対する検出力の感度を示している。

非農業者／非ばく露者の NHL 割合を 0.2387 と推定した場合、OR 1.2、1.5、2.0、3.0 を検出する検出力（片側）は、それぞれ 30.5%、81.5%、99.7%、99.9%以上となる。なお、Waddell ら（2001 年）は、マラチオンを使用した NHL 症例 91 例と使用しなかった非農家 243 例を対象に、OR は 1.6、95%CI は 1.2-2.2 と報告している。

以上のことから、真のオッズ比が 1.2、1.5、2.0、3.0 の場合に、影響量の拡大の差がどの程度存在するかをシミュレーションするために SAS が使用された。以下は、マラチオンと NHL の検出力解析のために SAS で作成された表で、SAS ベースのシミュレーション結果から、影響の拡大のマグニフィケーションの大きさを示している。

1.2、1.5、2.0 及び 3.0 の真のオッズ比を与えられた影響量の拡大を示す SAS シミュレーション結果 [(a)]

真値		N 個の解析データセット	Power[(b)]	観察された有意な OR の分布			
非ばく露群における罹患者の割合	OR			N	10%台	中央値 (%inflation)	90%台
0.1194 (1/2 background)	1.2	1,000	0.22	220	1.4	1.5 (25)	1.8
	1.5	1,000	0.66	661	1.5	1.7 (13)	2.0
	2	1,000	0.97	972	1.6	2.0 (0)	2.5
	3	1,000	1.0	1,000	2.4	3.0 (0)	3.7
0.2387 (19 background)	**1.2**	**1,000**	**0.32**	**323**	**1.3**	**1.4 (17)**	**1.6**
	1.5	1,000	0.81	812	1.4	1.6 (7)	1.8
	2	1,000	1.0	997	1.6	2.0 (0)	2.4
	3	1,000	1.0	1,000	2.5	3.0 (0)	3.6
0.4774 (29 background)	1.2	1,000	0.34	337	1.3	1.4 (17)	1.6
	1.5	1,000	0.87	872	1.3	1.5 (0)	1.8
	2	1,000	1.0	1,000	1.6	2.0 (0)	2.5
	3	1,000	1.0	1,000	2.4	3.0 (0)	3.7

ロジスティック回帰モデルが、2 つのグループのオッズ比を計算するために使用された。最尤推定値が存在しない場合（おそらくいずれかのグループの症例がゼロであったため）、EXACT 検定をいくつかのデータセットの解析に使用した。

(a)：片側検定、a = 0.05、非ばく露群=1,018、マラチオンばく露群=238、反復 N=1,000（データセット）。(b)：このシミュレーションから得られた検出力は近いかもしれないが、組み込みの SAS (PROC POWER)や Stata (power 2-proportion)のような統計ソフトウェアのプロシージャを使用している。これは、シミュレーションされたデータセットの数が十分でないためかもしれない。しかし、1,000 回の反復は、検出力を十分に推定し、統計的に有意な結果（ここでは、≤0.05）が得られた場合の影響量の拡大を説明するのに十分である

p<0.05 で統計的に有意な結果が得られた場合、影響量の中央値は、非ばく露群の NHL の割合と真のオッズ比（1.2 から 3.0 まで）によって、1.4 から 3 まで変化することに注意する。例えば，非農家の NHL の割合が 0.2387 の場合、真の OR が 1.2（表中の太字の行）とすると、統計的に有意な OR の半分は中央値 1.4 を超え、半分は下回っていることになる。さらに、統計的に有意な OR の大部分（90%）は 1.3 以上であることが観察され、少数（10%）は 1.6 以上であることさえ観察された。

まとめると、疫学研究の検出力は、規制当局や他の人がそのような研究を評価する際に考慮すべき重要な要素である。十分な検出力を持つ研究は、与えられた規模の真の効果が実際に存在する場合に検出する可能性が高くなるだけでなく（検出力の古典的な定義は、II 型エラーや偽陰性の問題に関連している）、影響が存在しないが（偶然にも）事前に選択された閾値（統計的有意性の 0.05 レベルなど）を超えた場合に、影響を拡大したり誇張したりする可能性が低くなるだろう。研究が適切な検出力を持っている場合（例えば、80%以上）、観察された影響量は真の影響量を再現する可能性が高く、この影響量の観察された偶然の変動は、未知の真の値を中心に対称的に分布を再現する。これらのシミュレーションと Ioannidis によるオリジナルの研究及び Gelman と Carlin（2014 年）による拡張研究から得られるメッセージは、研究は偽陰性（タイプ II エラー）を回避するために適切な検出力が必要であるだけでなく、統計的に有意な（または他の閾値を通過した）影響量のために影響量の拡大を回避するために適切な検出力が必要であるということである。Gelman と Carlin（2014 年）はさらに進んで、そのような「後ろ向き計画の計算は、統計的に有意でない影響量よりも統計的に有意な影響量に関連しているかもしれない。統計的に有意な結果の解釈は、基礎となる影響のもっともらしい規模によって大きく変わる可能性がある」と述べている。研究が適切な検出力を持っていれば体系的なリスクの膨張は皆無であるが、統計的に有意な効果をもたらす検出力不足の研究の影響推定値は、実質的なリスク膨張の可能性があり、その解釈は真の（基礎となる）影響の現実的な推定値に依存することに注意することである。

理想的には、公表されている文献研究は、検出力解析を実施し、文書化するべきである。それ以外にも、公表されている文献は、読者がこのような検出力計算（あるいは Gelman と Carlin（2014 年）が言うところの（後ろ向き）デザイン計算）を行うのに十分な情報を提供すべきである。上記の 2 つの例では、著者は読者に検出力と影響量の拡大を計算することができる十分な情報を提供していた。これは常にそうとは限らない。検出力の計算に使用された情報が文献に部分的にしか提供されていなかったり、提供された情報がそのような計算ができない方法で構成されていたりすることがある[43、44]。例えば、著者が三分位または四分位を決定するためにばく露量の代わりに症例数を使用している場合（これはグループ間での症例数が一定であることから証明される）、または著者が複数のがん症例をまとめてグループ化し、その数を使用して三分位を決定している場合、必要な入力が得られないため、ここに示されている検出力（またはデザイン）の計算は不可能である。集積及び報告される数値及びデータは、著者及び文献の間で必ずしも標準化されているわけではないので、一つの強い推奨事項は、文献が（補足的またはオンラインデータであっても）検出力を推定するために必要な情報を報告することを義務付けることであり、このような評価を査読者及び関心のある読者の両方ができるようにすることであろう。

以上の解析から、影響量の膨張現象の潜在的な意味合いは、疫学研究を評価する上で重要な考慮事項であることが示唆されたが、この現象に関するいくつかの注意点を覚えておくことが重要であり、疫学研究の解釈にどのように考

[43] 例えば、Stella Koutros ら（2012 年）の出版物「Risk of Total and Aggressive Prostate Cancer and Pesticide Use in the Agricultural Health Study」で発表されたマラチオンばく露と侵攻性前立腺がんの関連性のレビューでは、発表された論文に重要な情報が提供されていなかったため、パネルはマラチオンばく露群と非ばく露群の比較の力を算出できなかった。文献及び文献の補足文書から、非ばく露群の症例数を容易に把握することができたが（本文中の表 2）、非ばく露群または各ばく露量（四分位）の被験者数は入手できなかったようである。我々は、文献の補足文書の表 1 の情報から非ばく露者群の被験者数及び各四分位の被験者数を導入しようとしたが、表 1 の情報は、ばく露者を症例数の四分位に基づいてグループに分類するという他の多くの AHS の出版物と一致しない方法で示されていたため、それを行うことができなかった。

[44] 検出力の計算に使用された情報は、文献では部分的にしか提供されていないことがある。例えば、我々は Laura Beane-Freeman ら（2011）による AHS 研究文献「Atrazine and Cancer Incidence Among Pesticide Applicators in the Agricultural Health Study（1994-2007）」に記載されている情報から、様々な甲状腺がんの比較に関連した検出力を計算した。この文献では、著者らは被験者をばく露に基づいて四分位に分類するのではなく、その代わりに、すべてのがん症例を合わせた総症例数に基づいて被験者を分類またはグループ化した。このようにして、すべてのタイプのがんの症例数は、分類されたグループ間で同じであり、したがって、対象となる任意の特定のがん（例えば、甲状腺、ここでは）の症例数は、グループ間で同じではなく、被験者の数は、グループ間で同じではなかった。この例では、文献は、（i）参照 Q1：N=9,523、（ii）Q2、Q3 及び Q4 の総被験者数を提供した。N=26,834 人（表 1）及び（iii）参考 Q1 の甲状腺がん症例数=3 人（表 2）を提供した。しかし、比較群（Q2、Q3 または Q4）のそれぞれの正確な被験者数は得られなかった。

慮すべきであるかについても留意すべきである。

- 第一に、この現象は、対象となる影響が統計的（またはその他の）閾値を通過するような検出力不足の研究では影響量が膨張する傾向があるが、他にも推定値を帰無に向かって逆方向に偏らせるバイアスが存在する可能性がある。このバイアスは、影響量抑制と呼ばれることがある。おそらく、これらのバイアスの中で最もよく知られているのは、本文で議論されている非差異的誤分類バイアスである。これは、一般的に（常にではないが）帰無値に向かって予測可能なバイアスを生じさせ、それによって影響量を体系的に過小予測する。これが常に正しいとは限らず、（少なくとも小規模の低い検出力の研究については）影響量の拡大のような対抗要因や相殺要因が存在する可能性があることを認識することは、重要な前進である。特に、検出力不足の研究では、例えば、非差異的誤分類バイアスから生じるかもしれない帰無値へのバイアスを潜在的に相殺する（そして、多分相殺以上に）ことができる程度に、帰無値から離れた方向に偏った推定値をもたらすことがある。いずれにしても、重要なことは、統計的に有意な結果を得るための影響量の推定値に少なくともいくつかの最低限の一致度を持たせることができるようにするためには、十分な検出力のある研究が必要であるということを認識することである。
- 第二に、そして本文で述べられているように、影響量の拡大は、研究者（またはそのような研究を解釈する規制者）の側で、その研究が検出力不足の場合に、与えられた有意性の閾値（例：$p<0.05$）を通過するか、または一定の大きさ（例：OR>3）を達成する影響を特定することに焦点を当てた努力に関連している。この現象は、統計的有意性（または影響量）の「事前スクリーニング」を行う際に最も懸念される現象である。規制者や政策決定者などが、事前に決められた統計的閾値を「通過」した関連性のみに焦点を当てて行動することを避け、研究が検出力不足の場合に効果量の拡大を評価して判断するためにその影響量を使用する場合、この現象はあまり懸念されない。影響量の拡大の決定は、研究や研究デザインの機能や欠陥ではなく、**むしろその研究がユーザーコミュニティによってどのように解釈されるかの機能である**ことに注意する。

残念なことに、与えられた量よりも大きい、またはある統計的閾値を通過して「発見された」影響量に注目が集まる傾向が時々ある。これらの「発見」がどのように考慮されるべきかに関しては Ioannidis によって推奨されている（Ioannidis、2008 年）。

「最初に仮定された発見の時点では、影響量を判断することはおろか、関連性が全く存在するかどうかもわからないのが普通である。最初の原則として、影響量については慎重にならなければならない。不確実性は、単に CI（95%、99%、99.9%であるかどうかは関係ありません）だけでは伝わらない。

新たに提案された関連性については、提案された影響の信頼性と正確性はケースによって異なる。ヒトは次のような質問をするだろう：この分野の研究コミュニティは、研究成果を主張するために、広く統計的な有意性や同様の選択の閾値を採用しているのか？発見は小規模な研究から生まれたのか？分析に大きな変動の余地があるか？選択的な報告から保護されていないか（例：プロトコールが前もって完全に利用可能ではなかったか）？特定の「ポジティブな」結果を発見し、促進することに興味を持っている人や組織はあるか？最後に、影響を相殺する力は最小限に抑えられているか？

- 第三に、影響量の膨張現象は、発見科学全般に適用可能な一般的な原則であり、疫学の特異的悩みや弊害ではないことを覚えておくべきである（Ioannidis、2005 年；Lehrer、2010 年；Button、2013 年；Button ら、2013 年；Reinhart、2015 年）。先に示したように、これは薬理学、遺伝子研究、心理学研究、そして最も頻繁に引用される医学文献の多くでしばしば見られる。このような真実性の膨張は、研究の規模が小さく、検出力が不足している場合に起こり、そのような研究では結果に大きなばらつきがあるからである。これは、多くの研究者が同様の研究を行っており、「新しい」または「刺激的な」結果を発表するために競争している場合には特に問題となる（Reinhart、2015 年）。

まとめと結論

影響量の拡大または「真実の膨張」とは、影響を「発見」するために統計的またはその他の閾値を満たす必要がある

検出力不足の研究から得られた効果測定値の場合に、オッズ比、相対リスク、または率比の推定値が誇張されることがある現象である。この現象は、疫学や疫学研究に特有のものではなく、むしろ、影響が存在するかどうかを判断するために、影響量や統計的有意性に関連するような、研究の規模が小さく、事前に設定された閾値が使用される傾向にあるあらゆる科学に見られる。このように、疫学研究の利用者がこの問題とその潜在的な解釈の結果を認識することが重要である。特に、検出力不足の研究から発見された関連性は、統計的または他の同様の閾値を通過したことに基づいて強調されたり、注目されたりするが、それは帰無値から系統的にバイアスがかかっている。統計的閾値を通過した「発見された」関連性の結果として、特定の研究で観測された影響量が帰無値から遠ざかるかどうかはわからないが（非差異的な誤分類を示す特定の研究が必ずしも帰無値に向かって偏るとは言えないのと同じように）、ここで提示された説明と実行されたシミュレーションによって説明されているように、偶然がそのような偏りをある程度有利にすることはわかっている。別の言い方をすると、統計的またはその他の閾値を通過した影響量に基づいて影響量に注目したり、報告したり、行動したりすることを選択することで、バイアスが導入される（Yarkoni、2009 年）。繰り返しになるが、これは研究がユーザーによってどのように解釈されるかに関連する問題であり、研究デザインに内在するものでもなければ、優れた科学技術の原則や実践に関連するものでもない。

上記の問題に対する（部分的な）解決策の 1 つは、観察された影響量が帰無値から系統的に偏ってしまう可能性があることを認識した上で、事前に定められた閾値を通過した疫学研究の影響量または検出力不足の研究から生じた場合の統計的に有意な影響量を慎重に解釈することである。このようなアプローチでは、著者が研究の検出力を報告するか、著者が読者に十分な情報を提供する必要がある。検出力が実質的に 80%未満の研究から得られた影響量は、おそらく実質的に（特に検出力が 50%未満の場合）誇張する可能性があることを認識した上で、適切な程度の疑いを持って解釈する必要がある。この誇張の潜在的な程度は、対象となる健康影響のバックグラウンド率、研究のサンプルサイズ、対象となる影響量など、多くの問題に依存する。より具体的には、(a) 対象となる健康影響のバックグラウンド率が低い場合、(b) 研究のサンプルサイズが小さい場合、(c) 対象となる影響量が弱い場合、研究の検出力（その影響量を検出するための）は低く、統計的に有意な結果において影響量が誇張される傾向が高くなる。対象とする健康影響のバックグラウンド率が低い集団で、小さな、または弱い影響を調査する低検出力研究では、影響量の誇張が最も大きくなる傾向がある。その結果、PPR パネルは、疫学的文献にこのような計算を組み込むか、または読者が計算を実行できるような重要な情報を含めることを推奨している。具体的には以下の通り。

特定の農薬ばく露と疾病との間の関連が統計的に有意であることが判明した場合、特に（推定される）検出力の低い研究では、データ利用者は、統計的に有意な影響量推定値（OR または RR）がどの程度人工的に拡大、または誇張されているかを判断するために、様々な検出力計算（または検出力解析）を実行すべきである。これは、疫学研究で明確に報告される 3 つの値を必要とする。(i) 非ばく露群の被験者数（罹患者と非罹患者を含む）、(ii) ばく露群の被験者数（罹患者と非罹患者を含む）、(iii) 非ばく露群の罹患者数である。リスク管理者は、次に、ばく露群と非ばく露群の間の所定の（予め決められた）影響量の差を検出するために、対象となる目標値（典型的には、OR または RR）を選択でき、影響量の規模が、対象となる研究で推定された影響量をどの程度説明できるかを評価することができる。

(i) 多くの疫学研究はしばしば検出力不足であり、(ii) 著者が検出力の計算や（場合によっては）計算に必要な情報を文献の中で提供することは一般的ではなく、(iii) 影響量の拡大の現象は一般的に疫学分野ではほとんど認識されていないようであるため、上記の PPR パネルの勧告を実施するためには、研究者／助成機関、出版社、研究資金提供者の側での努力が必要である。上記のように、この分野での実践の現状は悲観的になるかもしれないことを示唆しているが、研究者である Kate Button（Button、2013 年）が Nature Reviews Neuroscience 誌に掲載したこのトピックに関するオピニオン・ピース（Button ら、2013 年）で、楽観について慎重な理由を提供した。

これらの問題に対する認識は高まっており、問題を認識することは、現在の実践を改善し、解決策を特定するための第一歩である。出版バイアスの問題を一朝一夕に解決するのは難しいが、研究者は確立された（しかし、しばしば無視される）科学技術の原則を採用することで、研究の信頼性を向上させることができる。また、研究者は、確立された（しかし無視されることが多い）科学技術の原則を採用することで、研究の有用性／信頼性を向上させることができる。

1）研究の計画や結果の解釈において、統計的な検出力を考慮する。

2）研究方法と結果の開示には誠実さを高める。

3）研究計画書、解析計画、さらにはデータまでもが公表されるようにする。

4）供給源を共有し、サンプルサイズと再現力を向上させるために協力して作業する

上記の一連の推奨事項と考えは、サンプルサイズと神経毒性学の背景で設定されているが、疫学を含むあらゆる発見科学に広く適用可能である。まとめると、疫学研究の実施と報告が公衆衛生に基づいた選択をする際に規制機関にとって有用なものとなるためには、改善の余地が大いにあるが、問題点はより明確にされ、認識され始めており、今後も楽観的に考えられる理由がある。

参考文献

Beane Freeman, LE, Rusiecki, JA, Hoppin, JA, Lubin, JH, Koutros, S, Andreotti, G, Hoar Zahm, S, Hines, CJ, Coble, JB, Barone Adesi, F, Sloan, J. Sandler, DP, Blair, A, and Alavanja, MCR. Atrazine and cancer incidence among pesticide applicators int eh agricultural health study (1994–2007). Environ Health Perspect, 119, 1253–1259.

Button K, 2013. Unreliable neuroscience? Why power matters. *The Guardian* newspaper (UK). 10 April 2013 Available online: https://www.theguardian.com/science/sifting-the-evidence/2013/apr/10/unreliable-neuroscie nce-power-matters [Accessed 6 September 2017]

Button K, Ioannidis JPA, Mokrysz C, Nosek BA, Flink J, Robinson ESJ and Munafo MR, 2013. Power failure: why small sample size undermines the reliability of neuroscience. Nature Reviews Neuroscience, 14, 365–376.

Cohen P and Chen S, 2010. How big is a big odds ratio: interpreting the magnitudes of odds ratios in epidemiological studies. Communications in Statistics: Simulation and Computation, 39, 860–864.

Gelman A and Carlin J, 2014. Beyond power calculations: assessing type S (sign) and type M (magnitude) errors. Perspectives on Psychological Science, 9, 641–651.

Ioannidis JP, 2005. Why most published research findings are false. PLoS Med, 2, e124.

Ioannidis JP, 2008. Why most discovered true associations are inflated. Epidemiology, 19, 640–648.

Jones RR, Barone-Adesi F, Koutros S, Lerro CC, Blair A, Lubin J, Heltshe SL, Hoppin JA, Alavanja MC and Beane Freeman LE. Incidence of solid tumours among pesticide applicators exposed to the organophosphate insecticide diazinon in the Agricultural Health Study: an updated analysis. Occupational and Environmental Medicine, 72, 496–503.

Koutros, S, Beane Freeman, LE, Lubin, JH, Heltshe, SL, Andreotti, G, Hughes-Barry, K, DelllaValle, CT, Hoppin, JA, Sandler, DP, Lynch, CF, Blair, A and Alavanja, MCR, 2013. Risk of total and aggressive prostate cancer and pesticide use in the agricultural health study. American Journal of Epidemiology, 177, 59–74.

Lehrer J, 2010. The truth wears off: is there something wrong with the scientific method. New Yorker. 13 December, 2010. Available online: http://www.newyorker.com/magazine/2010/12/13/the-truth-wears-off [Accessed September 2017]

Reinhart A, 2015. Statistics Done Wrong: the WOEfully complete guide. No Starch Press (San Francisco, CA).

Rosenthal JA, 1996. Qualitative descriptors of strength of association and effect size. Journal of Social Service Research, 21, 37–59.

Taubes G, 1995. Epidemiology faces its limits. Science, 269, 164–169.

Waddell BL, Zahm SH, Baris D, Weisenburger DD, Holmes F, Burmeister LF, Cantor KP and Blair A, 2001. Agricultural use of organophosphate pesticides and the risk of non-Hodgkin's lymphoma among male farmers (United States). Cancer Causes Control, 12, 509–517.

Wynder EL, 1997. Epidemiology Faces its Limits – Reply. Invited Commentary: Response to Science Article, "Epidemiology Faces Its Limits". American Journal of Epidemiology, 143, 747–749.

Yarkoni T, 2009. Ioannidis on effect size inflation, with guest appearance by Bozo the Clown. 21 November 2009. Available online: http://www.talyarkoni.org/blog/2009/11/21/ioannidis-on-effect-size-inflation-with-guest-appea rance-by-bozo-the-clown/ [Accessed on 6 September 2017]