SCIENTIFIC OPINION

EFSA Journal

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

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.

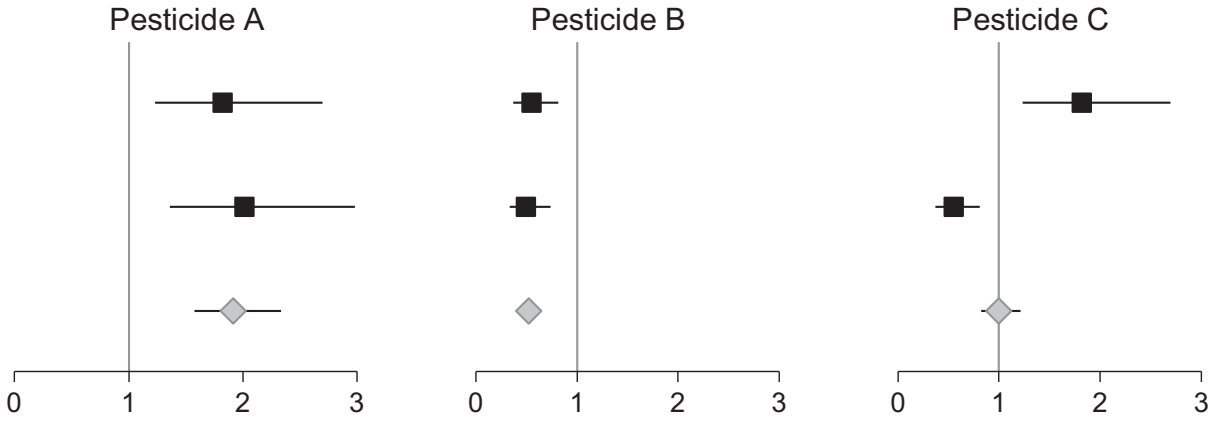


Figure 1: Forest plots from a fictitious example in which each of three pesticides (A, B and C) is evaluated in a meta-analysis of two studies. The x-axis in each plot represents the estimated risk ratio of the disease of interest comparing exposed and unexposed individuals. The squares denote the estimated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

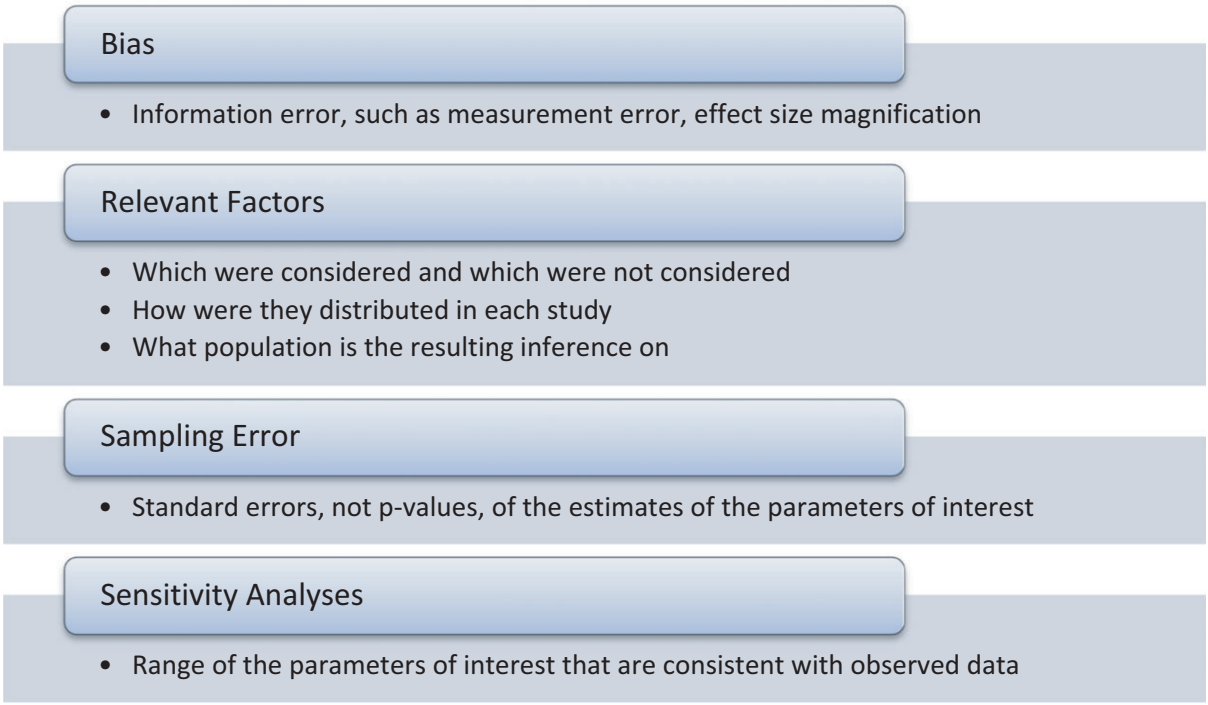


Figure 2: Items to consider when evaluating and comparing multiple studies

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.

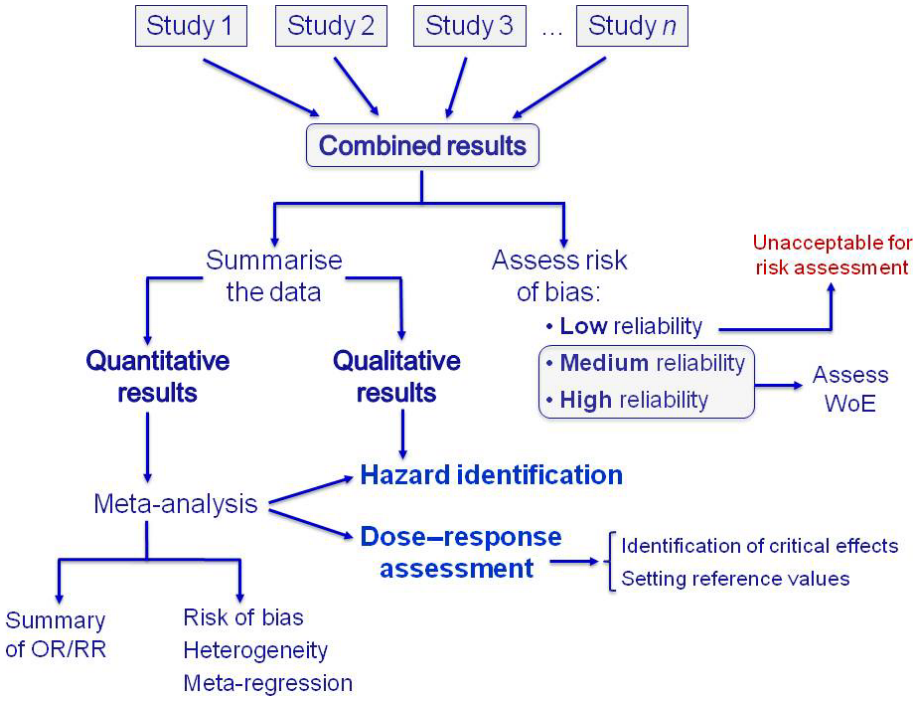


Figure 3: Methodology for utilisation of epidemiological studies for risk assessment

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.

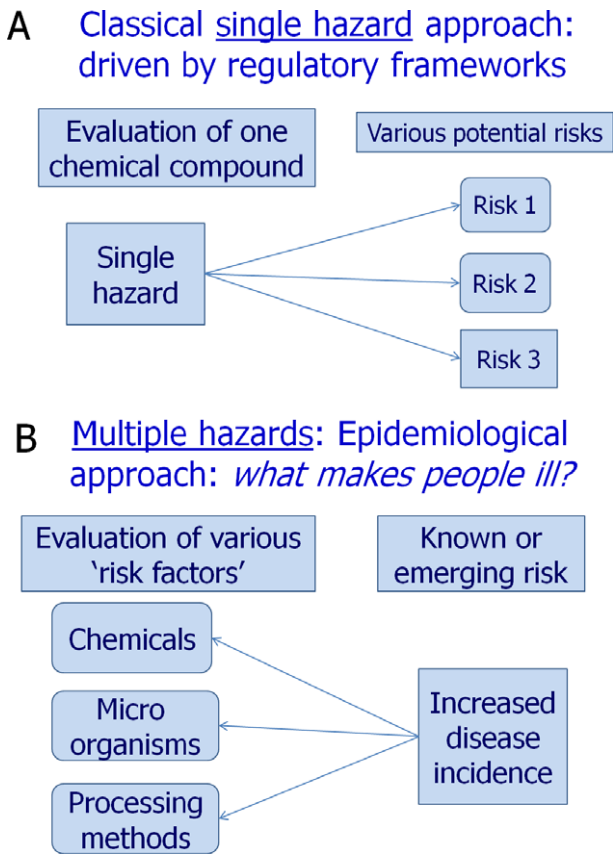


Figure 4: Role of epidemiological studies when compared to classical toxicological studies

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.

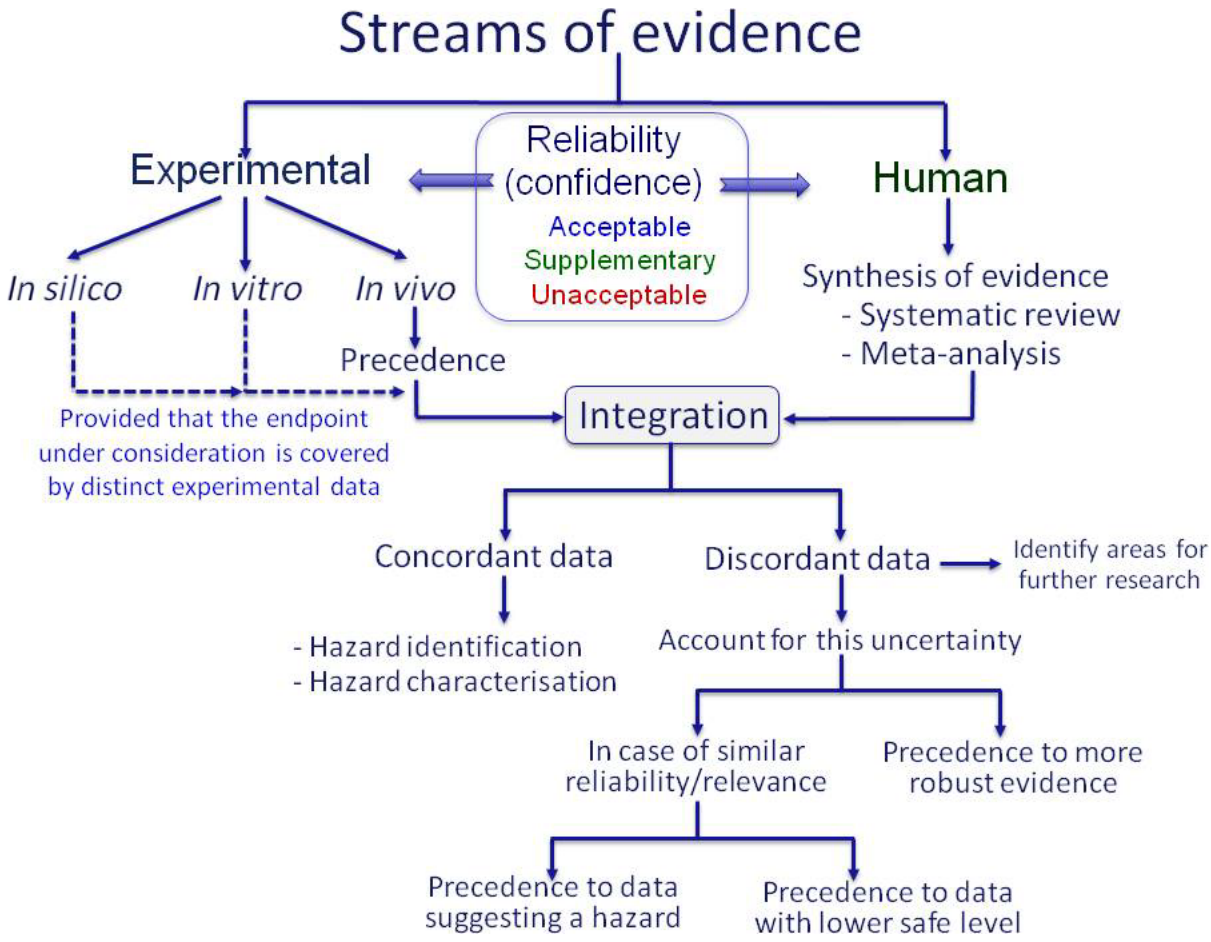


Figure 5: Methodology for the integration of human and animal data for risk assessment

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).

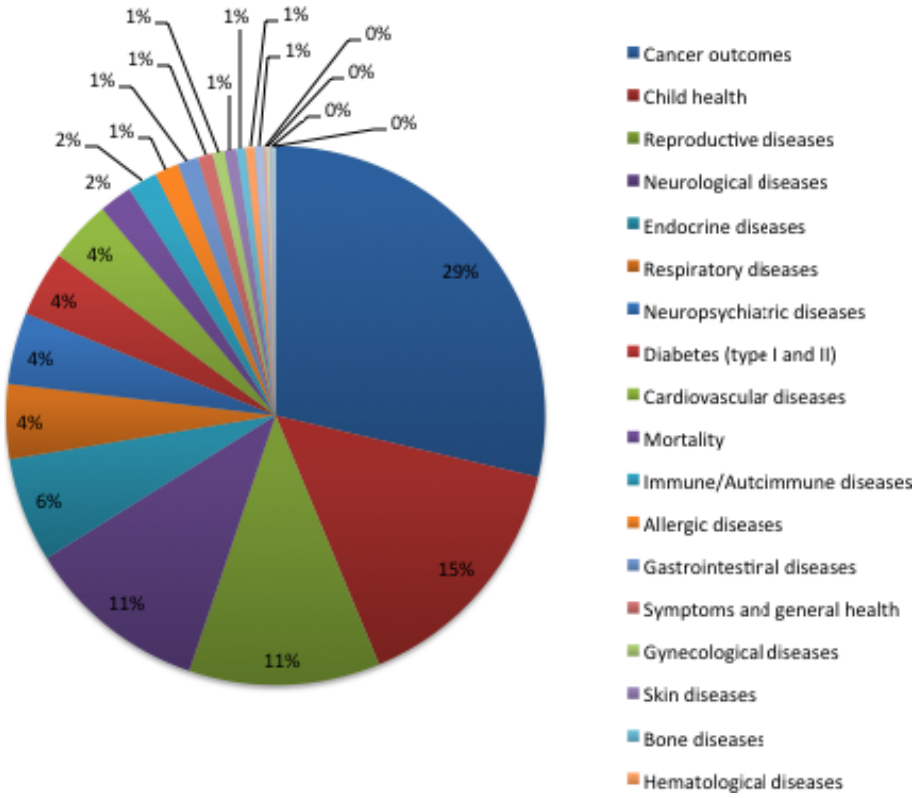


Figure A.1: Major outcome categories and corresponding percentage of studies examining those outcomes among the publications reviewed by the EFSA external scientific report (Ntzani et al., 2013)

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.

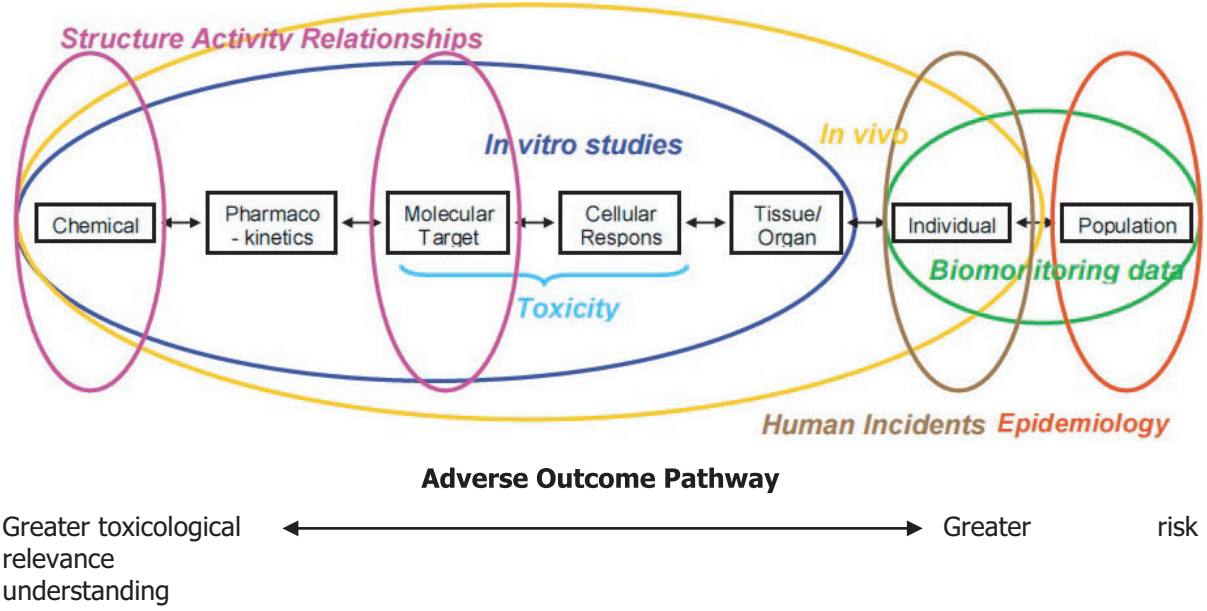


Figure C.1: Source to Outcome Pathway: Chemical effects across levels of biological organisation (adapted from NRC, 2007)

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)]

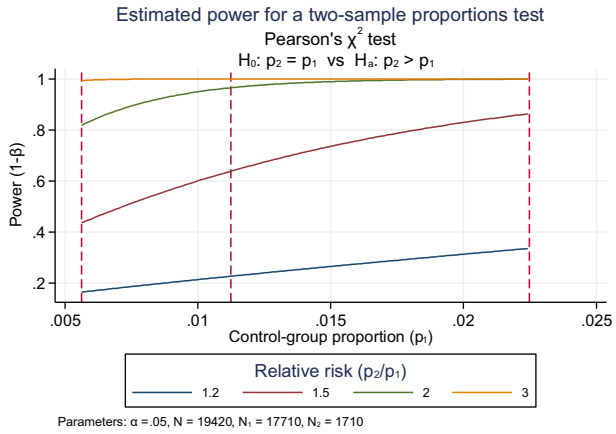
$N_{control}$	$N_{exposed}$	Proportion control[(b)]	Proportion exposed	Relative risk	Power
17,710	1,710	0.00562	0.00674	1.2	0.1634
17,710	1,710	0.00562	0.00843	1.5	0.4353
17,710	1,710	0.00562	0.01124	2.0	0.8182
17,710	1,710	0.00562	0.01685	3.0	0.9935
17,710	**1,710**	**0.01124**	**0.01348**	**1.2**	**0.2259**
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

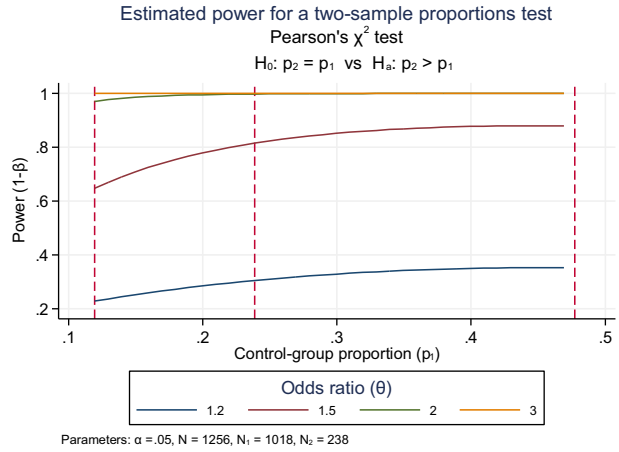
Stata code used to generate the above results: `power two-proportions ('= 0.5 * 243/1018' '= 243/1018' '= 2 * 243/1018'), test(chi2) OR (1.2 1.5 2.0 3.0) n1(1,018) n2(238) one-side table(N1:''N control'' N2:''N exposed'' p1:''proportion control'' p2''proportion exposed'' OR:''odds ratio'' power:''power'').`

(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]