EFSA's Draft 2014 Scientific Opinion on the risks to public health related to the presence of bisphenol-A (BPA) in foodstuffs: a critical appraisal

Paul Whaley
The Policy from Science Project
22 April 2014

Contents

Summary of findings

This critical appraisal identifies major methodological obstacles to accepting that EFSA's draft 2014 Opinion on BPA (EFSA 2014) is a valid, maximally-useful synthesis of the evidence of BPA's toxicity.

The appraisal has been conducted according to a set of general principles derived from the Cochrane Collaboration's approach to systematic reviews, developed to minimise bias and maximise reproducibility of reviews of medical trials. These principles are outlined in a previous report on the use of systematic review methods in chemical risk assessment (Whaley 2013).

Primary concerns relating to the validity and utility of the Opinion

1. The search and selection methods for finding evidence only retrieved a partial snapshot of the current literature relevant to evaluating risks to health posed by BPA. Documentation of search methods is insufficient to evaluate the extent to which this might have affected the results of the Opinion.

2. The appraisal of the methodological quality (reliability) of the literature in the hazard assessment of BPA is not valid, failing to in fact measure the methodological quality of research.

3. The weight-of-evidence analysis is insufficiently documented to allow its validity and consistency of conduct to be evaluated.

4. There appears to be two parallel, distinct risk assessment methodologies in the one document: the initial objective of characterizing risk is apparently superseded by a second process in which the Panel is concerned with identifying studies which would warrant a change to the current TDI for BPA. It is therefore ambiguous as to whether or not the Opinion fulfils its terms of reference.

Note on (4) EFSA could have done one of two things in developing the Opinion: either (a) find a study which overturns the TDI; or (b) do a full risk assessment synthesising all available data. Throughout the Opinion, the approach is consistently presented as if it is delivering (b).

However, the requirements described in the Hazard Characterisation (section 3.9.7) are consistent not with (b) but with (a): the Panel's interest is in whether a study is of sufficient quality and provides enough information to yield a dose-response curve judged by the Panel to be robust enough to be the basis of a risk assessment.

This TDI procedure differs fundamentally to the hazard and risk assessment procedure described in the Opinion, holding studies to a different set of inclusion criteria and different criteria for appraising methodological quality than those used in the hazard assessment process. This makes the hazard assessment redundant to the TDI procedure: little of the hazard assessment procedure is relevant to conducting the TDI procedure.

The problem is that the Opinion presents the redundant hazard assessment process as if it is in fact pivotal to the risk assessment, thereby presenting its conclusions as if they are based on one process when in fact they are based on a quite different and independent process. This is likely to be at the very least confusing, if not outright misleading, for users of the Opinion.

Discernible improvements relative to previous Opinions on BPA

The extra documentation provided in this Draft Opinion compares favourably with previous Opinions, particularly in the level of detail given in the appraisal of the methodological quality of studies included in the review.

However, this improvement in transparency also reveals a lack of clarity with regard to best practice in finding, selecting, appraising and synthesising evidence of the toxicity of BPA before interpreting this into a risk assessment.

This should not be a surprise: best practices in these areas are not yet established. EFSA's commitment to transparency in documentation and improving risk assessment methodology is therefore a useful contribution to strengthening the scientific foundations of chemical risk assessment.

Recommendations for the 2014 Opinion

1. In consultation with stakeholders, choose a single objective to either
   a. evaluate existing studies to determine if there is a published study which would force revision of the current TDI for BPA (does not require a hazard assessment), or
   b. conduct a full risk assessment based on synthesis of all available data (does require a hazard assessment).

The following steps should be taken regardless of a choice between (a) or (b):

2. Review all the data relevant to the risk assessment of BPA within a specified time scale, either cutting off at 2012 or including all 2013 data.

3. Fully document the results of the search process and list the excluded studies, giving the primary reason for exclusion of each.

If (b), the subsequent steps should be taken:

4. Evaluate each relevant study on its own merits rather than treating previous Opinions as definitive appraisals of the state of the evidence relating to the toxicity and risk assessment of BPA.

5. Present results for the appraisal of the reliability and weight of studies on a study-by-study basis (as done in the neurotoxicity lines of evidence) rather than grouping several studies into single judgments of reliability and weight.

6. For appraising the reliability of studies, use a tool which better measures the credibility of a study than the guidance currently given in the document.

7. Fully document the method for (a) interpreting study reliability into weight of evidence, and then (b) interpreting weight of evidence into likelihood of hazard, ensuring that the same method is followed for each end-point.

Summary table

EFSA Panel on Food Contact Materials, Enzymes, Flavourings and Processing Aids. (2014) **Draft Scientific Opinion on the risks to public health related to the presence of bisphenol A (BPA) in foodstuffs.** European Food Safety Authority, Parma, Italy.

Question	Appraisal
1. Does the review address a clearly-focused question?	**Unclear.** The stated objective of the Opinion is to collate the published evidence of hazards to health posed by BPA and interpret this into a dose-based risk assessment. However, the hazard characterization section appears to redirect the purpose of the Opinion towards determining whether or not any individual study exists which warrants changing the tolerable daily intake (TDI) for BPA. These are not the same exercise and how they relate to each other is unclear. The overall objective of the Opinion is therefore unclear.
2. Has the review been conducted in accordance with procedures defined in a pre-published protocol?	**Unsatisfactory.** There was no pre-published protocol according to which the Opinion was conducted.
3. Is there a comprehensive declaration of interests and contributions?	**Unclear.** Declarations of interests of the Panel and Working Group members have to be downloaded from the EFSA website. How these interests may have influenced the Opinion has to be inferred from this documentation. This is hampered by neither interests of specific relevance to the Opinion being identified, nor any account of what each person contributed to the development of the Opinion. It is not therefore possible to determine how specific interests of Panel and Working Group members might have influenced the various elements of the Opinion.
4. Did the authors locate all the research which might have been relevant to answering the specified question?	**Unsatisfactory.** The Opinion does not appear to have taken into consideration all the research which was relevant to its Objective: a search of the literature quickly finds papers from 2010-2012 which appear to be relevant yet are not referenced in the document; in addition, the Opinion openly acknowledges that the search method for papers published after 2012 is partial.
5. Did the authors use a screening process which put forward to analysis all the research relevant to the question?	**Unsatisfactory.** EFSA used a different selection process for including studies from 2013 (expert judgment of relevance to the review) than it did for studies from 2010-2012 (a specified list of inclusion criteria). The role of allowing TDI calculation as an inclusion criterion is unclear.
6. Did the authors do enough to assess the external validity (relevance) of the included studies?	**Unclear.** No explicit methodology for appraising the directness of evidence provided by a study is given. Although directness of evidence is sometimes described in the appraisal of research, it is not clear what the criteria for relevance are, nor how they affect the weight given to a study in the overall analysis, nor if they are consistently applied.
7. Did the authors do enough to assess the internal validity (reliability) of the included studies?	**Unsatisfactory.** There are three issues in the appraisal of the reliability of studies: the criteria for methodological quality are not valid; there is insufficient evidence that the criteria are consistently applied; and it is not clear how fulfilment of the criteria translates into a judgment of reliability.
8. Did the authors combine the results, directness and methodological quality of evidence into a statement of size and strength of evidence for an effect?	**Unclear.** The validity of the Panel's method for synthesizing data in the Opinion in the weight-of-evidence analysis cannot be evaluated due to insufficiency of documentation. The approach taken to synthesis is inconsistent between health end-points, with neurotoxicity and metabolic end-points taking a different approach in comparison to the other end-points of interest. The validity of using older Opinions as start-points for the hazard assessment is questionable.
9. Is there a clear answer to the review?	**Satisfactory.** The summary section appears to provide an accurate summary of the major findings of the Opinion.

The Appraisal

EFSA Panel on Food Contact Materials, Enzymes, Flavourings and Processing Aids (2014)

Draft Scientific Opinion on the risks to public health related to the presence of bisphenol A (BPA) in foodstuffs

European Food Safety Authority, Parma, Italy

METHOD

This critical appraisal evaluates the utility, reproducibility and validity of EFSA's 2014 Draft Opinion on the risks to health posed by bisphenol-A (EFSA 2014).

1. **Utility:** does the approach taken in the Opinion yield an answer of maximum use to stakeholders in the risk assessment and risk management of BPA?

2. **Validity:** do the procedures used in the Opinion to evaluate and combine evidence about risks to health of BPA yield an accurate picture of what is actually known about these risks?

The general principles are derived from the Cochrane Collaboration's approach to systematic reviews, developed to minimise bias in reviews of medical trials. These are described in the Policy from Science Project report on how systematic review techniques might apply to chemical risk assessment (Whaley 2013).

The appraisal is divided into nine components, each being an essential part of a useful and valid synthesis of evidence. Analysis of each of the nine components results in a judgement of satisfactoriness:

- **Satisfactory:** The component is conducted according (within reason) to a clear, valid and consistent procedure

- **Unclear:** There is insufficient documentation to allow evaluation of the component

- **Unsatisfactory:** There is positive evidence of inconsistent or invalid procedure in the conduct of the component

This appraisal focuses on the methodology of the review rather than specific issues of whether or not the appraisal of quality of a particular study is correct. These are not entirely independent; comment is therefore made when material content of the review affects the appraisal of its validity or utility.

Declaration of Interests and Contributions

Paul Whaley developed the scheme for analysing the Opinion, conducted the analysis and wrote this report. Yannick Vicaire and Dr Crispin Halsall reviewed this report and the analytical scheme.

Paul Whaley Paid work over the last 5 years relevant to this report: ChemSec (researcher); Cancer Prevention and Education Society (retainer as Scientific Advisor, editor and producer of Health & Environment); Standardisation Network for Sustainability (Technical Committee Expert Representative); Centre for Sustainable Healthcare (researcher, writer); BetterValueHealthcare (web developer, writer); New Harbour (researcher); Health Care Without Harm (employee); Réseau Environnement Santé (researcher, project manager). He is undertaking a part-time PhD at Lancaster Environment Centre on applying systematic review methods to chemical risk where he is provided with office space and equipment, a doctoral training grant, a waiver on fees for first year of study, and a grant from Radical Futures in Social Sciences. No investments or other financial interests to declare. No Board memberships.

Yannick Vicaire Is a Board Member of four NGOs concerned with environmental issues: the Health and Environment Alliance; Greenpeace Czech Republic; Centre National d'Information Indépendante sur les Déchets; and InterreAction. In the last 5 years he has been paid for consultancy work by: Greenpeace International; Comité Catholique contre la Faim et pour le Développement; Sherpa; and Ethique sur l'étiquette. Since 2008, he has been employed by: Greenpeace France; Agir pour l'Environnement; and Réseau Environnement Santé. He has grants or grants pending from: the European Environment and Health Initiative; Fondation pour un terre humaine; Conseil Régional Rhône-Alpes; and Fédération Nationale de la Mutualité Française.

Dr Crispin Halsall CChem MRSC. Reader at Lancaster Environment Centre, Lancaster University. PhD supervisor to Paul Whaley.

APPRAISAL

1. Objective of the Opinion

Does the review address a clearly-focused, relevant question?

Unclear. The stated objective of the Opinion is to collate the published evidence of hazards to health posed by BPA and interpret this into a dose-based risk assessment. However, the hazard characterization section appears to redirect the purpose of the Opinion towards determining whether or not any individual study exists which warrants changing the tolerable daily intake (TDI) for BPA. These are not the same exercise and how they relate to each other is unclear. The overall objective of the Opinion is therefore unclear.

Explanation. EFSA states that the objective of the draft Opinion is to "evaluate the toxicity of BPA for humans, including for specific (vulnerable) groups of the population (e.g. pregnant women, infants and children, etc.) and considering all relevant toxicological information available" and "characterise the human health risks taking into account specific groups of the population" (p2 and p16).

In broad terms, it appears that the objective of the Opinion is therefore to: identify the likely hazards to health posed by BPA, as evidenced by the current literature; then determine whether current exposures to BPA mean any of the likely hazards pose an actual risk to health.

However, the process of hazard characterization (section 3.9.7) seems to move the goalposts. Here, EFSA states: "A prerequisite for the risk characterisation step is hazard characterisation, involving examination of a possible dose-response relationship for the effect under consideration and identification of a dose level at which the effect is not anticipated to occur (NOAEL) or a dose level at which the incidence of the effect is considered to be low (LOAEL or BMDL)."

This statement is the reason for exclusion of the 12 studies which resulted in identification of BPA as a hazard to the mammary gland: "The CEF Panel considered that none of these studies [identifying BPA as a hazard in the mammary gland] were sufficiently robust methodologically or showed a consistent dose-response to be used as the basis of a revised TDI."

The 2013 U.S. FDA/NCTR study is considered by EFSA "to be a detailed and methodologically robust study … that could be used on its own for risk assessment purposes" and was therefore analysed to see if it would yield a NOAEL or LOAEL. Of this, "the Panel concluded that the data could not be used to provide such a BMDL, since the outcome of modelling contained considerable uncertainty."

Consequently, it appears there are two apparent objectives in the Opinion. The first is described in the terms of reference of the Opinion, setting the objective as a broad appraisal of all the relevant literature into a statement of the risks posed to health by BPA. This appears to be later superseded by a second objective, apparent from the hazard characterization, to identify any study which is suitable to changing the current TDI for BPA.

Given there are two objectives in play, one of which appears to supersede the other, and that the relationship between the two is not clearly documented, it is unclear as to what the Objective of the Opinion actually is. Consequently, it is unclear whether or not the Opinion fulfils the terms of reference provided by EFSA.

Note on low-dose effects The relatively brief consideration given to low-dose effects and non-monotonic dose response curves (NMDRCs) in the risk assessment of BPA (section 1.3, page 24), in justifying their exclusion from the risk characterization of BPA, may be overly perfunctory relative to the importance of the issue to the risk assessment of BPA and the level of stakeholder interest in this issue.

2. Adherence to a pre-defined protocol

Has the review been conducted according to procedures defined in a pre-published protocol?

Unsatisfactory. There was no pre-published protocol according to which the Opinion was conducted.

Why this matters. The publication of protocols prior to the conduct of reviews is important for reducing the impact of author biases (such as expectation bias) and improves transparency of methods. In this case, pre-publication and peer-review of the Panel's methodology could have enabled identification of a number of the methodological concerns identified in this document prior to publication of the Draft Opinion, facilitating a more efficient and robust review process.

3. Declaration of authors' interests

Is there a comprehensive declaration of interests and contributions?

Unclear. Declarations of interests of the Panel and Working Group members have to be downloaded from the EFSA website. How these interests may have influenced the Opinion has to be inferred from this documentation. This is hampered by neither interests of specific relevance to the Opinion being identified, nor any account of what each person contributed to the development of the Opinion. It is not therefore possible to determine how specific interests of Panel and Working Group members might have influenced the various elements of the Opinion.

Explanation Information about the interests of authors, be they financial, academic or otherwise, is important for allowing users to put the findings of a review in their full context. In order to do this, users also need to be informed of what each author contributed to a review. The potential influence of interests on the formulation of the Opinion would be more transparent if solely the relevant interests were presented to the user and this were given in the Opinion document along with information about contributions.

4. Comprehensiveness of search strategy

Did the authors locate all the research which might have been relevant to answering the specified question?

Unsatisfactory. The Opinion does not appear to have taken into consideration all the research which was relevant to its Objective: a search of the literature quickly finds papers from 2010-2012 which appear to be relevant yet are not referenced in the document; in addition, the Opinion openly acknowledges that the search method for papers published after 2012 is partial.

Explanation. Searching the literature for papers about BPA toxicity published between 2010 and 2013 readily returns studies which are not referenced in the Opinion. The search method is also inconsistent, with studies published after 2012 being found not by a systematic search method but on the basis of "expert judgement" (p199). That EFSA is aware of this, stating that "the Panel acknowledges that these studies may not represent the entire body of evidence that has become available between January 2013 and the date of endorsement of this Scientific Opinion" (p199) does not make the method any less partial.

Examples of studies published 2010-2013 of possible relevance to the Opinion which are not referenced in the Opinion documentation

- Effects of prenatal and postnatal exposure to a low dose of bisphenol A on behaviour and memory in rats. Gonçalves, Carjone Rosa; Cunha, Raquel Wigg; Barros, Daniela Marti; Martínez, Pablo Elías. *Environmental Toxicology and Pharmacology*, 2010, Vol.30(2), pp.195-201
- Anxiety- and Depressive-Like Behaviours in CD-1 Mice Developmentally Exposed to Bisphenol A. Nelms, J; Ward, M; Meyer, A; Miller, M; Sable, H. *Neurotoxicology And Teratology*, 2011, Vol.33(4), pp.509-509
- The Effects of Maternal Exposure to Bisphenol A on Allergic Lung Inflammation into Adulthood. Bauer, Stephen M; Roy, Anirban; Emo, Jason; Chapman, Timothy J; Georas, Steve N; Lawrence, B. Paige. *Toxicological Sciences*, 2012, Vol. 130(1), pp.82-93
- Developmental exposure to bisphenol A leads to cardiometabolic dysfunction in adult mouse offspring. Cagampang, F. R; Torrens, C; Anthony, F. W; Hanson, M. A. *Journal of Developmental Origins of Health and Disease*, 2012, Vol.3(4), pp.287-292
- The impact of neonatal bisphenol-A exposure on sexually dimorphic hypothalamic nuclei in the female rat. Adewale, HB; Todd, KL; Mickens, JA; Patisaul, HB. *Neurotoxicology*, 2011, Vol.32(1), pp.38-49
- Corticosterone-regulated actions in the rat brain are affected by perinatal exposure to low dose of bisphenol A. Poimenova A, Markaki E, Rahiotis C, Kitraki E. *Neuroscience*. 2010 May 19;167(3):741-9. doi: 10.1016/j.neuroscience.2010.02.051. Epub 2010 Feb 26.

- Pubertal exposure to bisphenol A disrupts behaviour in adult C57BL/6J mice. Yu C, Tai F, Song Z, Wu R, Zhang X, He F. Environ Toxicol Pharmacol. 2011 Jan;31(1):88-99. doi: 10.1016/j.etap.2010.09.009. Epub 2010 Sep 15. PubMed PMID: 21787673.

According to the Opinion, EFSA's search method yielded between 115 and 3731 citations per database (p198), yet there are only approximately 450 references in the Opinion (pp165-194). Studies of potential relevance to the Opinion have therefore either not been located or have been excluded from the review. How these omissions or exclusions might have affected the findings of the Opinion cannot be evaluated because there is insufficient documentation of the search and inclusion processes.

5. Selection of evidence for review

Did the authors employ a screening process which selected for analysis all the studies of actual relevance to their research objective?

Unsatisfactory. EFSA used a different selection process for including studies from 2013 (expert judgment of relevance to the review) than it did for studies from 2010-2012 (a specified list of inclusion criteria). The role of allowing TDI calculation as an inclusion criterion is unclear.

Explanation. Studies published in 2013 and later are included in the Opinion on the basis of "expert judgement" i.e. depending upon whether or not the Panel thought they should be included. This is a different process than applied to studies published between 2010-2012, where studies were included on the basis of meeting the set of criteria specified on page 196.

The use of pre-specified inclusion criteria instead of expert judgement in the process of selecting studies is important because it reduces the risk of error in the overall review, by ensuring that all studies of actual relevance are included and not just the studies which the experts believe to be relevant (in other words, it ensures selection bias does not distort the evidence base of the review).

As is the case for the search component of the Opinion, that the Panel "acknowledges that the studies selected from the publications in 2013 may not represent the entire body of relevant evidence published up to the date of the launch of the public consultation of this opinion" (page 199) does not make this selection method any less prone to error. The consistent, valid approach would be to either exclude all data from 2013 or use the same selection procedure for 2013 studies as for studies from 2010-2012.

Role of TDI in selection In section 3.9.7 the Opinion states that, in order for a study to be included in the Hazard Characterisation, it must permit a BMDL, NOAEL or LOAEL to be calculated. In relation to this the Opinion goes on to say: "the Panel considered that none of these studies were sufficiently robust methodologically or showed a consistent dose-response that could be used to compare with the current TDI or as the basis of a revised TDI."

This seems to introduce a criterion of methodological quality in relation to calculating a TDI as an inclusion criterion for this part of the risk assessment procedure, suddenly excluding the data on which the hazard assessment is based and changing the apparent objective of the Opinion (as discussed above for component 1). What is going on procedurally at this point of the Opinion is very unclear.

6. Appraisal of directness of evidence

Did the authors apply a fair test of *external validity* (relevance) to each of the studies included in their review?

Unclear. No explicit methodology for appraising the directness of evidence provided by a study is given. Although directness of evidence is sometimes described in the appraisal of research, it is not clear what the criteria for relevance are, nor how they affect the weight given to a study in the overall analysis, nor if they are consistently applied.

Explanation. Because it is not ethical to conduct toxicological research on humans, the toxicity of chemicals to humans has to be inferred from epidemiological, animal and in-vitro studies. Depending on the type of research (such as from a monkey model to a human in comparison to a mouse model to a human, or observation of a marker of an effect to an actual toxic effect itself), these inferences can be more or less direct.

The more direct a model is, the more weight it should carry in the analysis. There should therefore be a scheme for consistent appraisal of the directness (the external validity) of the evidence from a study, so this can be consistently factored into how much emphasis is placed on its results in the final analyses. Since EFSA does not provide such a scheme, it is not possible to evaluate whether or not the methodology here is valid or consistently applied.

7. Appraisal of methodological quality of evidence

Did the authors apply a fair test of *internal validity* (methodological quality or reliability) to each of the studies included in their review?

Unsatisfactory. There are three issues in the appraisal of the reliability of studies: the criteria for methodological quality are not valid; there is insufficient evidence that the criteria are consistently applied; and it is not clear how fulfilment of the criteria translates into a judgment of reliability.

Explanation. Studies obviously vary greatly in terms of methodological quality (their internal validity, or "reliability" in the vocabulary of the Opinion), with better research more likely to home in on the true effect that a chemical has on health. The better a study's methodology (the greater its internal validity), the more weight it should carry in the analysis. This means studies need to be appraised for methodological quality according to a consistent standard which actually measures the likelihood that their results reflect the true toxicity of the chemicals they are testing.

EFSA describes the criteria it uses for appraising methodological quality of studies in Appendix 1 of the Opinion, with specific criteria listed in tables 24 and 25. Although this presentation is welcome, as things stand it seems unlikely that the tool can reliably distinguish better studies from worse:

- the criteria measure features such as consistency between studies, reporting quality and conformity with guidelines, none of which are valid markers of the methodological quality of studies;
- documentation is insufficient to show that the criteria are being consistently applied;
- the extrapolation of the appraisals of methodological quality into judgements of reliability of the research are insufficiently documented and of questionable validity.

Challenges to the validity of EFSA's tool for appraising methodological quality of evidence

EFSA's tool is only valid if the criteria which it uses to measure methodological quality actually measure methodological quality. Although some criteria do this, not all do so. Furthermore, some criteria which are important for appraising methodological quality are absent from the tool. (Note that the following list of problems with the Panel's approach is not exhaustive.)

The Opinion measures quantity of information instead of methodological quality of research. Single dose studies are considered of worse quality than a multi-dose study. Although single-dose studies provide less information than multi-dose studies they do not necessarily provide inaccurate information.

The Opinion measures consistency between studies instead of methodological quality. It is not clear what the Panel means by "consistency of results" when judging study plausibility. Two interpretations suggest themselves; neither one tracks the methodological quality of a study.

- *Agreement in outcome does not necessarily indicate higher methodological quality.* If three studies using different methods disagree about an outcome, it cannot be inferred that the studies are methodologically weaker than they would have been had they agreed. For example, one of the studies could be methodologically robust while the other two suffer substantial weaknesses which make them unimportant in understanding the toxicity of the compound in question. In this case, it would be incorrect to conclude that the strong study is somehow less credible because of the limitations of the weaker studies. Alternatively, each study could suffer from weaknesses which render them equally biased. That they agree in their results would not make them any more likely to be true.
- *Disagreement in outcomes does not necessarily indicate lower methodological quality.* The Panel may be concerned about studies which use the same method and yield different results i.e. the precision of the methodology. Precision is not, however, the same thing as methodological quality: many imprecise studies can still home in on the true effect size, while precise studies can be biased. Although precision needs to be taken into account when estimating effect size from a group of studies, it is not appropriate to use it as a criterion of methodological quality.

The Opinion measures reporting quality instead of methodological quality. Many of the criteria are based on reported methodological quality rather than actual methodological quality. This is explicit for the epidemiology tool (p203) and in the requirements for reporting strain, age, body-weight etc. of animals. Deliberate downgrading of studies on the basis of weaknesses in reporting is an error because it conflates what researchers actually did with what they say they did, when the relationship between reporting quality and accuracy of results is in fact unclear. Basing judgements of methodological quality on reporting quality is explicitly advised against in the Cochrane Collaboration tool for appraising risk of bias in research (Higgins et al. 2011).

The Opinion measures conformity with guidelines instead of methodological quality. That a study is performed according to guidelines does not necessarily mean it is methodologically superior to a study which is not performed according to guidelines. This is because it is possible for a non-guideline study to be of equivalent quality to a guideline study. Use of this criterion is doubly problematic because it will systematically downgrade all the non-guideline research with which the Panel is concerned, regardless of the actual methodological quality of the research.

The Opinion fails to measure some aspects of methodological quality. Some criteria for appraising methodological quality are absent from the Opinion. Allocation concealment, random allocation to test groups, attrition of test subjects and selective reporting are important considerations in appraising the methodological quality of a piece of research, are either absent or incompletely described.

Did EFSA appraise all the evidence according to the same quality criteria?

There is insufficient evidence that these criteria were applied consistently by the Panel, with only partial reporting of weaknesses identified in studies and no systematic record of how each study performed against the tool (only selected strengths and weaknesses are presented in Appendix II).

How did EFSA extrapolate judgements of reliability of lines of evidence from fulfilment of quality criteria?

Unclear as to how satisfying reliability criteria is interpreted into an overall judgement of reliability. There is not enough information about how the satisfaction of quality criteria yields a judgement of reliability. In this respect, the Opinion should include guidance as to the likelihood and extent to which an identified weakness leads to overestimating or underestimating effect size (the risk of bias introduced by a weakness) in a study.

Issue with the use of a reliability scale. The Panel combines judgements of methodological weaknesses and strengths into an overall score for reliability of a study or group of studies. This score is a scale. For a scale to be valid, each score has to be a greater or lesser quantity of the same identified quality (for example, colour saturation can be scored on a scale but colour hue cannot). Although methodological quality might intuitively seem like the sort of thing which can be scored, quality of research is determined by a varied mix of non-comparable elements, making it invalid to combine them. The Cochrane Collaboration explicitly advises against the use of scales and scores in describing the methodological quality of research (Higgins et al. 2011).

8. Synthesis of the evidence

Did the authors combine the results, directness and methodological quality of evidence into a statement of size and strength of evidence for an effect?

Unclear. The validity of the Panel's method for synthesizing data in the Opinion in the weight-of-evidence analysis cannot be evaluated due to insufficiency of documentation. The approach taken to synthesis is inconsistent between health end-points, with neurotoxicity and metabolic end-points taking a different approach in comparison to the other end-points of interest. The validity of using older Opinions as start-points for the hazard assessment is questionable.

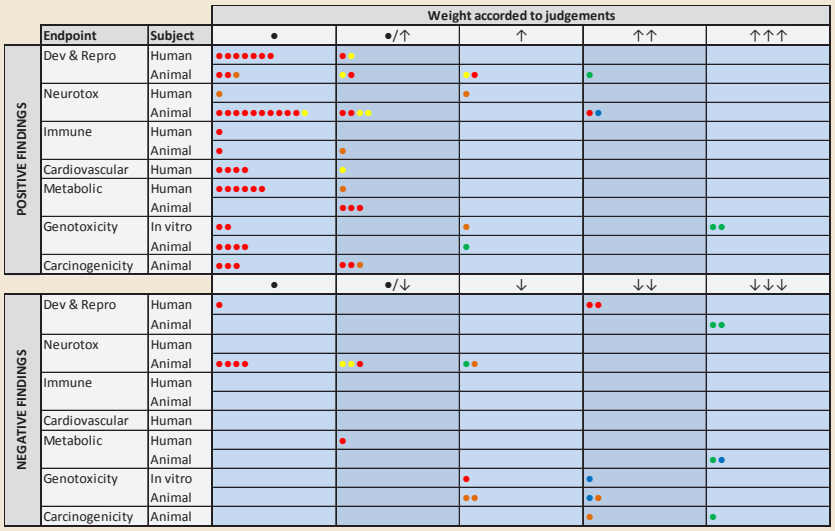
Explanation. The purpose of reviewing research is to produce: a better estimate of the effects on health which a chemical might have than can be given by any single study; the strength of the evidence supporting that estimate; and a view of what is and is not known in relation to these effects, thereby placing the estimate in its full research context. Overall, there is insufficient documentation of the process by which the Panel synthesized the data it reviewed into its overall statement of results to judge its validity.

Specific issues with synthesis of data in the Opinion

Interpretation of reliability into weight of evidence. Figure 1 shows how judgments of reliability on their own are poor predictors of how much weight a piece of evidence carries in the overall analysis. For example, two judgments of low reliability carry a relatively large amount of weight against the likelihood that BPA is a developmental toxicant, while one judgment of high reliability carries relatively little weight against the neurotoxicity of BPA.

This inconsistency could be explained by whatever factors the Panel is considering when interpreting reliability into weight; however, these factors are not documented so the weighting process cannot be evaluated for validity.

An additional obstacle to evaluating the validity of judgments of reliability and weight concerns how several studies are often combined into single reliability ratings and judgments of weight. It is not possible then to determine whether or not each individual study has been validly appraised. Only the neurotoxicity evaluation consistently appraised studies individually; this is the most transparent approach and should be taken for all end-points in the Opinion.



Endpoint	Subject	Judgment count (not countable)
Dev & Repro	Human	12 (1)
	Animal	10 (3)
Neurotox	Human	2 (1)
	Animal	26 (5)
Immune	Human	1 (1)
	Animal	2 (1)
Cardiovascular	Human	5 (0)
Metabolic	Human	8 (3)
	Animal	5 (7)
Genotoxicity	In vitro	7 (0)
	Animal	9 (1)
Carcinogenicity	Animal	8 (3)

Key (1 dot = 1 judgement of reliability)	
●	Judged to be of high reliability
●	Judged to be of med/high reliability
●	Judged to be of medium reliability
●	Judged to be of med/low reliability
●	Judged to be of low reliability

Figure 1. Summary of reliability and weight judgments in EFSA's Draft 2014 Opinion on BPA. Each judgment made by the Panel is indicated as a dot. The colour of the dot corresponds to the reliability of the judgement. The position of the judgement in the columns corresponds to the weight given to the judgement in the analysis of the likelihood of the hazards BPA may pose to human health. Judgements which combined reliability such that they could not be represented in the table (such as simultaneous up and down arrows, or of being low and high reliability) were excluded from the table as not countable. Note that it is judgments, not studies, which are represented in the tables, as some judgments are the combined reliability of several studies.

Validity of using old Opinions as start-points for hazard assessment. Using older Opinions as discrete start-points for current risk assessment is methodologically dubious because they will have been developed according to different methods to the current opinion, which means not all evidence in the final review will have been treated according to the same standards. It also runs the risk of propagating prior error into the current analysis. Given that the issues previously raised with regard to methodological weaknesses of EFSA's 2010 Opinion on BPA have not been addressed (Whaley 2013), it seems a full reappraisal of all data is warranted.

Consistency of approach between hazard end-points. The way that metabolic function is assessed as a hazard end-point seems inconsistent with the rest of the hazard assessment. The approach for non-metabolic end-points is to treat sub-end-points (the individual questions) as individual hazards which stand in their own right in the hazard assessment.

However, for metabolic hazards, the sub-end-points are given individual judgments of likelihood which are then combined into an overall judgment of likelihood. The result is that one hazard judged as "likely" (glucose or insulin resistance and pancreatic effects) disappears from analysis before the hazard characterization (p477).

Interpretation of weight into likelihood of hazard. There is insufficient documentation of how judgments of weight are interpreted into likelihood of hazard.

9. Answering the question

Is there a clear answer to the review which is representative of its main findings?

Satisfactory. The summary section appears to provide an accurate summary of the major findings of the Opinion.

Summary of Policy from Science Project analyses of EFSA Opinions on BPA

	2010 Opinion	2013 Draft Exposure Assessment	2014 Draft Opinion
Objective	●	●	●
Protocol	●	●	●
Interests	●	●	●
Search Method	●	●	●
Study Selection	●	●	●
External Validity	●	●	●
Internal Validity	●	●	●
Synthesis	■	●	●
Answer	●	●	●

KEY:	
●	**Satisfactory** The component is conducted according (within reason) to a clear, valid and consistent procedure
●	**Unclear** There is insufficient documentation to allow evaluation of the component
●	**Unsatisfactory** There is positive evidence of inconsistent or invalid procedure in the conduct of the component
■	**Component not appraised**

Readers should note that the appraisal of declaration of interests and contributions for the 2010 and 2013 Opinions has been changed from "unsatisfactory" (as reported in Whaley 2013) to "unclear" after revision of the criteria for judging the validity and utility of approach to each component.

Bibliography

EFSA Panel on Food Contact Materials, Enzymes, Flavourings and Processing Aids. (2014) *Draft Scientific Opinion on the risks to public health related to the presence of bisphenol A (BPA) in foodstuffs.* European Food Safety Authority, Parma, Italy.

Higgins, J. P. T.; Altman, D. G.; Gotzsche, P. C.; Juni, P.; Moher, D.; Oxman, A. D. et al. (2011): The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. In *BMJ* 343 (oct18 2), pp. d5928. DOI: 10.1136/bmj.d5928.

Whaley, P. (2013): *Systematic review and the future of evidence in chemicals policy.* Policy from Science Project. Brussels, Belgium.