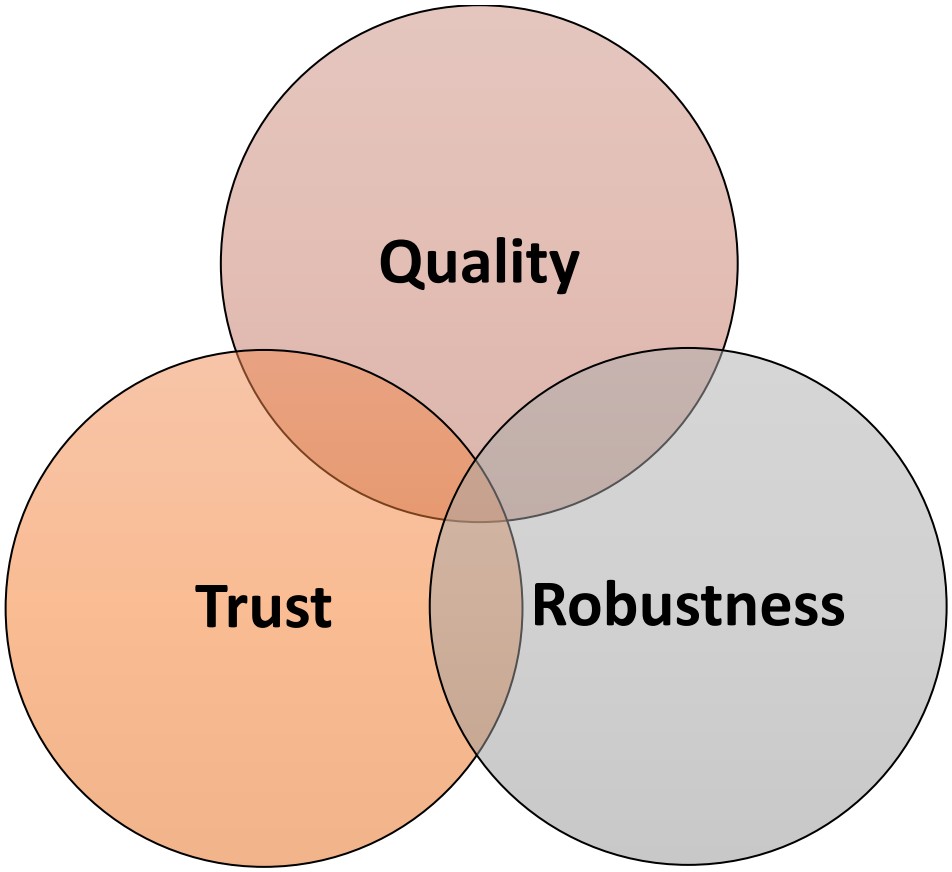
The FSA Science Council's Rapid Evidence Review on the Critical Appraisal of Third-Party Evidence

Final Report and Recommendations



Science Council Leadership Team: Prof Sandy Thomas (Science Council Chair), Prof Peter Gregory, Prof Sarah O'Brien

Secretariat: Dr Catriona McCallion, Dr Ben Goodall, Dr Chun-Han Chan, Mr Paul A. Nunn

DOI: https://doi.org/10.46756/sci.fsa.elm525

Contents

Suggested Citation: Food Standards Agency Science Council. 2021. Rapid Evidence Review on the Critical Appraisal of Third-Party Evidence. Report to the Food Standards Agency. https://doi.org/10.46756/sci.fsa.elm525

1 Summary of Principles

The Food Standards Agency (FSA) always seeks to ensure that its recommendations are made on the best-available evidence.

Following a request from the FSA Chair, the Science Council have sought to provide a framework that can guide those seeking to submit uncommissioned evidence to the FSA on its scientific principles and standards.

The Science Councils proposed framework is based on the principles of quality, trust and robustness.

By being transparent about the FSA's minimal expectations, we aim to help those who wish to submit evidence, typically in an effort to fill a perceived evidence gap or change a relevant policy or legislation. The framework also seeks to provides assurance to others on the processes in place within the FSA to assess evidence it receives.

When the FSA receives evidence, it will:

- be transparent about how the evidence is assessed and used to develop its evidence base, policy recommendations and risk communication;
- assess evidence in its proper context using the principles of quality, trust and robustness;
- seek to minimise bias in its assessments of evidence by using professional protocols, its SACs, peer review and/or multi-disciplinary teams;
- be open and transparent about the conclusions it has reached about any evidence submitted to it.

2 Recommendations

The Science Council recommends the FSA:

1. Adopt a framework for the assessment of uncommissioned third-party evidence, comprising:

    - a statement outlining how evidence is assessed by the FSA when submitted outside of standardised processes;
    - high-level principles and guidelines around the key themes of quality**,** trust, and robustness, to support the FSA's assessment of third-party evidence and provide third parties with an understanding of the standard of evidence expected by the FSA.

2. Review the adoption and integration of the framework after 12 months of use.

3 Introduction

In March 2020 then Chair of the FSA, Heather Hancock, asked the Science Council for its guidance on quality and assurance thresholds for third-party evidence, with the question:

> How should the FSA evaluate the robustness of evidence submitted by non-commissioned third parties in an effort to change our policy, in order to ensure that the evidence considered to inform our advice and recommendations is sufficiently robust and based on the most up to date scientific information?

In response to this question, the Science Council undertook a "Rapid Evidence Review" that has sought to provide a framework for the critical appraisal of third-party evidence. The aim of the framework is to act as a guide on the FSA's scientific principles and standards to those who wish to submit evidence to the FSA. The framework also provides assurance to others on the processes in place within the FSA to assess evidence it receives.

The Terms of Reference were agreed in September 2020 and preliminary recommendations were delivered to the FSA Chair and the Chief Executive in January 2021, in line with the end of the transition period for the UK's departure from the European Union. This final report was submitted to the FSA in June 2021, following a public consultation on the Science Council's proposed framework.

3.1 Scope

The Science Council seek to provide assurance to the FSA's assessment of the standard of evidence submitted unsolicited by third parties.

Uncommissioned third-party evidence may be submitted by a range of stakeholders including members of the public, industry representatives, consumer groups and others. It can be shared with a variety of motivations and aims, typically seeking to fill a perceived gap in knowledge or suggesting a change relevant to a policy or legislation.

If an organisation wishes to submit evidence related to placing a regulated product on the market, there is specific FSA guidance on the procedures and requirements for this. Regulated product authorisation was out of the Science Council's Rapid Evidence Review scope.

3.2 Approach and Structure of the Report

The Science Council's Rapid Evidence Review was conducted using four interlinked Work Packages (WPs) (Table 1).

Table 1 - Overview of the work packages constituting the Rapid Evidence Review on Third-Party Evidence

Work Package 1, Sept-Nov 2020 Complete
Desk study including: • Interviews with FSA and Food Standards Scotland leads to understand current practice in the assessment of evidence • Review of current best practice from international guidance and the wider literature on assessing evidence quality
Work Package 2, Oct-Dec 2020 Complete
FSA Scientific Advisory Committee (SAC) consultation: • Written responses from SAC members and discussion with SAC Chairs on the concept of good evidence • Engage as part of the internal consultation on an early draft of the proposed framework during Dec 20
Work Package 3, Nov-Dec 2020 Complete
Consultation with other government departments: • With support of the Chief Scientific Advisors Network, seek experience across government on the use of evidence assessment frameworks
Work Package 4, March-April 2021 Complete
Public consultation on the Science Council's proposed framework • One-month consultation supported by an online survey on the Science Council website.

3.2.1 Work Package 1

WP1 consisted of a desk study comprising a document review and a set of interviews with FSA and Food Standards Scotland (FSS) officials (Table 2).

The document review focused on the guidelines and standards published by international bodies of which the UK is or was a member, guidelines produced by the FSA itself, and outputs by national accreditation bodies. Additional literature reviewed was sourced from the reference lists of initial documents, a general web search and recommendations from the interviewed FSA officials. This exercise ensured that the primary aim of the Science Council, to provide a framework for the assessment of uncommissioned third-party evidence, made use of the good practice already in place.

The interviews were led by the Science Council members and Chair, Professors Peter Gregory, Sarah O'Brien and Sandy Thomas. Discussions were tailored and adapted to the interviewees area of expertise, with core emphasis on:

- how third-party evidence arrives at the FSA and how it is then handled within and between teams;
- how officials discern what good evidence is;
- the use of existing guidance and frameworks for assessing evidence.

Table 2 – list of interviews conducted with FSA and FSS lead officials as part of Work Package 1

Date (MM/DD/YY)	**Official(s) Interviewed**	**Position(s) Held**
10/2/2020	Amie Adkin	Head of Risk Assessment
10/5/2020	Paul Tossell	Policy Team Leader
10/15/2020	Paul Cook & Anthony Wilson	Microbiological Risk Assessment Team Leaders
10/12/2020	David Gott	Head of Toxicology
10/19/2020	Vanna Nasser-Aldin	Head of Analytics
10/20/2020	Joanne Edge and Frances Hill	Science Governance and Regulated Products Team Leaders
10/29/2020	Michelle Patel	Head of Social Science
11/19/2020	Steve Wearne	Director of Global Affairs
11/30/2020	Jacqui McElhiney, Marianne James and wider team	Head of Science, and Head of Risk Assessment; Food Standards Scotland

3.2.2 Work Packages 2 and 3

As part of WPs 2 and 3, the Science Council engaged with the FSA's Scientific Advisory Committees (SACs) and other government departments (OGDs) to gain insight into how they assess evidence and the use of any frameworks or guidelines.

Engagement with SACs was pursued through a call for information and a roundtable discussion at the SAC Chairs' meeting on 21 October 2020. Key individuals in OGDs were contacted through the Chief Scientific Advisor's network.

3.2.3 Work Package 4

WP4 was a public consultation on the Science Council's proposed framework for the critical appraisal of third-party evidence. By making a draft available for public

consultation, the Science Council aimed to ensure maximum clarity and usefulness for those who may submit evidence to the FSA. The consultation lasted between midday 22 March 2021 and 23:45 22 April 2021 and received eleven responses. A full report on the response to the public consultation is available on the Science Council Website.

4 The Science Council's Proposed Framework for the Assessment of Third-Party Evidence

FSA officials and FSA SAC members provided clear ideas of what constitutes good evidence, and highlighted frameworks and grading systems that are used in specific disciplines or areas of expertise; these are referenced in the following sections. They included guidance for professions such as the Government Analysis Function, which provides leading resources for analysts and social scientists working across UK Government [1].

While use of discipline-specific frameworks for the assessment of evidence is widespread, no overarching checklist or grading system has proven robust enough to remain an ongoing FSA component of practice. This may be because in some instances such tools did not provide enough flexibility given the breadth of evidence seen by the FSA, while in others, a grading system was not specific enough.

A single overarching checklist or grading system may also constrain critical thinking, which is essential for the effective assessment of evidence. In addition to the use of best practice guidance such as that produced by the Codex Alimentarius Commission, World Health Organisation (WHO) and the FSA itself, staff are trained using on-the-job coaching, mentoring and shadowing to impart good practice. Finally, in certain instances it may be legislation that governs the judgement of whether evidence is sufficient to be used for a food safety decision.

When OGDs and the FSA SACs were consulted, several SACs shared guidelines developed for specific areas of work and several OGDs also had frameworks for how internal evidence and analysis should be generated. The Department for International Development's[1] note on assessing the strength of evidence is a broadly applicable framework that is especially useful for the social sciences, but it states reviewers should think carefully about how exactly to apply the principles depending on the nature of the study.

The Science Council concluded that the principles and guidelines it developed should be high-level, encompassing the themes of **quality**, **trust** and **robustness,** which came to the fore during WP1-3. However, the Science Council note that these principles and the concepts underpinning them are not independent entities. For

[1] The Department for International Development was combined into the Foreign, Commonwealth and Development Office in 2020. This note will be referred to using "DfID" as it was produced before this change was made.

example, reliability is considered under the principle of quality and reproducibility is considered under robustness, but these concepts are inherently related and will impact each other.

The Science Council's framework is intended to encourage and facilitate the continued use of international best practice, guidance and standards that are appropriate to the disciplines and areas to which any uncommissioned evidence is related. Further, they aim to be flexible enough to allow reviewer discretion when considering evidence.

4.1 Principles and Guidelines

The Science Council have produced the statement below, outlining how the FSA should consider evidence, reflecting best practice from within the Agency. This accords with the FSA's established approach to science: that policies, decisions and advice will be made on the best available scientific evidence and analysis [2].

FSA statement on the consideration of evidence

When the FSA receives evidence, it will:

- be transparent about how the evidence is assessed and used to develop its evidence base, policy recommendations and risk communication;
- assess evidence in its proper context using the principles of quality, trust and robustness;
- seek to minimise bias in its assessments of evidence by using professional protocols, its SACs, peer review and/or multi-disciplinary teams;
- be open and transparent about the conclusions it has reached about any evidence submitted to it.

4.1.1 Quality

Evidence should be reliable and relevant to the question at hand. Clearly defining the context of the study and the question originally asked can help to identify if the evidence is relevant. Using well-recognised methods and data analysis can help to ensure it is both relevant and reliable. If a novel method is used, a clear explanation of why it has been used and what advantage it brings is important. Data and analysis should be clearly presented, with a narrative that directly links them to the conclusions within the study.

Clarity

- All evidence sent to the FSA should be clearly laid out, outlining the study approach, the data collected, and analysis performed.
- If evidence has been collated from several sources this should be clearly indicated, and the method used for its collation and integration described.

- Precise language should be used to describe the aims of the study or research question, relating this to the study design and conclusions.
- Methods should be described in enough detail that they could be independently reproduced – including the controls, reference standards and quality assurance measures used. This includes both study methods and methods for data analysis.
- A clear statement should be provided describing how data[2] were cleaned[3], processed and analysed, and why such approaches were taken. Providing the underlying data wherever possible, provides opportunity for its independent assessment, which can improve confidence in the conclusions reached.
- The conclusions of a study must be based on the evidence presented, with a clear narrative linking the data and analysis to those conclusions.

Relevance

- To assess the relevance of a study to a particular issue, the FSA will look at the context of the original study and the question(s) it was designed to answer. As key information about the way the study was conducted will be used to assess this, the clarity and transparency of the evidence are therefore important.
- The study design and the methods used should be justified with reference to the original question or hypothesis – including how potentially confounding variables[4] were controlled for.
- Consider the relevance of the study population, specimen or substance to the target population, specimen or substance. This is important when considering the biological relevance of a study and its conclusions.[5]
- If the study is qualitative, a comprehensive description of the context of the work should be included. For example, the culture, livelihood, community, socio-economic status and environment of participants.
- Statistical analysis is essential in scientific studies. Studies should include a clear outline of the methods used, and why they were chosen, with an explanation of what question the analysis aimed to answer. Statistical point estimates and confidence intervals are recommended alongside significance testing.

[2] We define data as all direct outputs from a study, including both quantitative and qualitative results, and digital images used to support analysis and conclusions.
[3] Data cleaning refers to the detection and removal of incorrect or corrupt data points, duplicates or empty fields, and ensuring consistency of units and formatting.
[4] Any additional variable that influences both the supposed cause (independent variable) and supposed effect (dependent variable).
[5] The work on Biological Relevance and Statistical Significance led by the Committees on Toxicity, Carcinogenicity and Mutagenicity will explore this in further detail.

- Where the evidence relates to a new method, outline the context in which the method should be used and why. Where relevant, make clear the advantages and drawbacks relative to validated methods.

Reliability

- Where possible, methods recommended by national and international bodies such as the Food and Agriculture Organisation of the United Nations (FAO), the Organisation for Economic Co-operation and Development (OECD) and the Codex Alimentarius Commission, and recognised by national and international standardisation bodies should be used.
- Good governance should be practiced when performing research. Refer to best practice guidelines such as the OECD's Principles of Good Laboratory Practice.
- Whether routine or not, all methods used should be referenced. If a standard method has been adapted, the study should state why and describe the differences. If a new method is proposed, a description of how it differs from the standard method(s), and where possible a comparative study should be provided.
- All evidence must include consideration of uncertainty[6]. Where possible this should be quantified using recognised methods. If the uncertainty is associated with an expert judgement or population sample, state whether it is qualitative or quantitative, and how it was discerned.
- Variability must also be considered, and where possible quantified[7]. Where variability has been controlled for in a study, consider if this affects generalisability to the target population, specimen, or substance.
- If mathematical models are used, the results of the sensitivity analysis performed should be provided, stating which parameters were tested, which were not and why

4.1.2 Trust

Transparency and impartiality are key in providing confidence that evidence is trustworthy. Evidence that is shared transparently will include access to underlying data, a clear explanation of the methods used and why, and the limits to the evidence provided. This includes stating uncertainties, variability and assumptions, indicating where results differ from comparable investigations and where there is dissenting opinion among experts. Any evidence and its assessment are at risk of

[6] We use the definition of uncertainty provided by the Committee on Toxicity: as the estimated sum of the limits in knowledge. We include limitations to apparatus, experimental techniques, models and study designs, as well as essential unpredictability.

[7] Variability is defined as the inherent heterogeneity between individuals or groups, or over time or space.

bias, but this can be mitigated by ensuring that sources of bias are recognised, peer review is performed, and challenge is built into the assessment process.

Transparency

- Openness and transparency are core principles of the way the FSA works; evidence submitted to the FSA should also demonstrate these principles as far as reasonably possible.
- In addition to clearly presenting all relevant data and associated analysis, access to the raw and omitted data from the study, including negative results, should be provided wherever possible. If this is not possible, state why. The FSA acknowledges that there may be legitimate commercial confidences and will respect these as far as reasonably possible.
- Known gaps in the evidence should be stated and limitations to models or study designs outlined. This includes assumptions on what is or is not important for the question being asked, and therefore what has been included or excluded from the study or model design.
- Consider alternative hypotheses and make comparisons to the published body of research on the area, stating where results differ or where there is disagreement in expert opinion.
- Clearly indicate when evidence is compiled from a range of sources. Reference all sources and state the method used to compile the evidence, for example, using widely accepted guidelines for evidence synthesis such as meta-analysis and systematic review procedures.

Impartiality and bias

- Increased risk of bias reduces the confidence in the outputs of a piece of evidence.
- All potential sources of bias should be clearly described, considering each stage of the study and any actions taken to mitigate them should be stated. The sources of bias and appropriate mitigating actions will be dependent upon on the type of study being performed.
- Where data are omitted from a study report, or where analysis is restricted to one or more subsets of the data available, this should be stated, with reasoning provided. If evidence is from a range of sources, the way in which sources were chosen or omitted should be given. This is consistent with the provision and reference to underlying data.
- Where expert judgement is used, state why, how the experts were chosen and the initial question that was asked of them. Any underlying data or evidence that the judgement is based upon should be provided, and a statement of uncertainty should be included with the judgement.
- If the evidence used is not published in a peer-reviewed journal, any independent critical review that has been performed should be described.
- In all instances, sources of funding and conflicts of interest must be stated.

4.1.3 Robustness

For evidence to be robust, a broad body of evidence should be considered from several perspectives, with each piece of evidence weighed based on its quality and trustworthiness. The body of evidence will be made more robust if the pieces of evidence are reproducible using the same and different methods. If an outcome is consistently observed when tested using different methods and populations, this provides confidence that the outcome itself is robust.

Consistency

- Describe how tests were replicated and the extent of any variation in the observed results.
- The clarity and transparency of a study, as well as the use of standard methods, reference standards and quality control methods can help ensure that a study can be repeated by other researchers.
- If several independent studies are performed repeating the same or similar tests and gaining the same or similar outcomes, this will increase confidence in the outcome.
- The robustness of an outcome can be tested by varying parameters within the study, and by using different methods to test the same relationship or outcome (triangulation). This may be done in a single study, or by comparing the outcomes of several studies.

Adequacy

- Explain the importance of the evidence with reference to the broader body of evidence to which it contributes. Consider whether evidence highlights any gaps in the existing body of evidence, and how much it increases the understanding of a new or emerging area.
- Different types of evidence may need to be combined for a comprehensive assessment of an issue to be undertaken[8]. Consider the other types of evidence that are required when assessing an issue and explain how your evidence relates to them.
- The adequacy of a piece of evidence will vary depending on the type of study and the question being asked. However, criteria such as the magnitude of any effect, the power of a study, and its applicability to the target population, specimen or substance may be considered.
- Significance testing is often used to indicate the magnitude of a result, but it is not by itself sufficient to indicate that a piece of evidence is strong or will translate to an important real-world impact. Consider the relevance of the

[8] For instance, as described in the Committees on Toxicity and Carcinogenicity's guidelines on the synthesis and integration of epidemiological and toxicological evidence.

study and the statistical test to the decision or policy that the evidence is being used to address.

5 Existing FSA Processes

The FSA prides itself in the transparent use of science and evidence to inform its advice and recommendations. Evidence enters the FSA from a variety of sources. For example, the FSA conducts and funds its own research to gain new insights; whilst food business operators, members of the public and those who represent them, may also submit evidence and views, commonly through consultations.

When new advice on food safety is required, the FSA's Risk Analysis Process will be used to assess the risk and advise on its handling. There are three components to the Risk Analysis Process:

Risk assessment involves using a scientific approach to identify and define hazards, and to estimate potential risk to human and/or animal health. This includes evaluating the likely exposure to risks from food and other sources.

Risk management is the consideration of potential measures to either prevent or control the risk. It takes into account risk assessment and consumers' wider interests in food to formulate a response.

Risk communication is the exchange of information and opinions throughout the risk analysis process. This can be between risk assessors, risk managers, consumers, industry, the academic community and any other interested parties. The FSA's Risk Communication Toolkit considers the most effective ways to communicate risk to consumers.

Uncommissioned evidence submitted by third parties to the FSA, may lead to the triggering of the Risk Analysis Process; however, this will depend on the standard of the evidence, as explored in Section 6, the hazard identified and the existing body of evidence that the FSA has available.

5.1 Other Legitimate Factors

The scientific process of understanding the risk a food poses to human health is at the heart of the Risk Analysis Process and providing consumers confidence that the food we eat is safe. However, there are considerations outside of the human or animal health risk assessment that are nevertheless essential for informing policy. These wider interests and considerations are referred to as "other legitimate factors" (OLFs) and are considered within risk management. The relevant legitimate factors considered will vary depending on the food or feed safety risk, but may include:

- Wider consumer interests such as environmental sustainability, animal welfare and food security;
- Consumer habits, perceptions, acceptability and preferences;
- Economic impacts, including who will bear the costs and who will benefit;
- Technical and feasibility considerations, including the ability to enforce and verify controls.

While the focus of this report has primarily been on evidence related to human health risk, if a third party submits uncommissioned evidence regarding an OLF that is scientific in nature, the considerations in this report will be applicable, and should be considered alongside further domain specific expert guidance, for example, that of the FSA's Advisory Committee for Social Science.

5.2 Communication

Throughout the interviews, communication was highlighted as essential for the appropriate handling of third-party evidence. The importance of communication in existing FSA processes is recognised within the Risk Analysis Process.

When the FSA receives uncommissioned third-party evidence, risk managers decide whether to seek an expert view from within the FSA (e.g. from Risk Assessors). Strong lines of communication between different teams are important for facilitating this.

6 What is Good Evidence?

This section is drawn from WPs 1-3. It sets out the existing national and international guidance and recommendations on best practice, and the wider academic and grey literature on how to assess evidence. It articulates how evidence is currently assessed in the FSA – and the expertise and judgement of officials and SAC members is therefore referred to throughout. The outputs presented here have informed the Science Council's recommended principles and guidelines on the appraisal of uncommissioned third-party evidence.

6.1 Guidelines and frameworks

Several relevant bodies also provide guidelines for assessing the quality of evidence and producing high quality evidence. For example, the European Food Safety Authority (EFSA) highlight relevance, reliability and consistency as key features for assessing the quality of evidence [3] and the UK Statistics Authority organise their Code of Practice for Statistics around the pillars of Value, Quality and Trust [4]. In 2014, the Department for International Development (DfID) outlined seven principles for high quality research: Conceptual Framing, Transparency, Appropriateness, Cultural Sensitivity, Validity, Reliability and Cogency.

Public Health England's Scientific Advisory Committee on Nutrition (SACN) provides guidelines for considering the plausibility of a conclusion [5], while within the FSA detailed guidance tailored to specific disciplines has been developed by SACs such as the Committee on Toxicity's (COT) in their Report on Synthesising Epidemiological Evidence [6]. The National Institute for Health and Care Excellence (NICE) provide checklists for assessing the quality of evidence from different types of study [7]. In many instances similar overarching themes are highlighted. Often these guidelines are targeted to commissioned or actively gathered evidence, rather than uncommissioned evidence; however, their themes remain applicable.

The following sections have been structured around the themes of **Quality**, **Trust** and **Robustness**, which the Science Council has agreed to frame its high-level principles and guidelines.

6.2 Quality

The quality of a piece of evidence is considered in this report using the criteria of clarity, reliability and relevance – that is, how the science was performed and communicated, and how it can be applied. Reliability and relevance can be used to assess the internal and external validity of a study, where internal validity considers whether a study was designed and performed appropriately given the question it was designed to answer; while external validity is concerned with whether the outputs can be generalised or reasonably applied to the issue at hand i.e. real-world relevance. The clarity of the study facilitates this assessment and the overall usefulness of the evidence.

6.2.1 Clarity

During interviews, respondents stated that the clarity of a piece of evidence is important when considering whether it is considered 'good' evidence. If information is difficult to extract, its significance can be lost. DfID highlighted "cogency" as key for assessing evidence; that studies should provide a clear, logical thread that runs throughout the entire report or submission, with the conclusions of the study clearly linked to the data and analysis presented [8].

The use of standard templates may assist the clear presentation of a study and can help ensure that all the required data is presented. EFSA provide a detailed template for reporting studies in their Guidance for Statistical Reporting [9]; Beronius et al. provide a template for reporting non-standard in vivo toxicology test results [10]; and EQUATOR (Enhancing the QUAlity and Transparency Of health Research) provides links to various reporting templates for health based research [11], including CONSORT (CONsolidated Standards of Reporting Trials) which provides a template for reporting randomised controlled trials (RCTs) [12], STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) which provides a template for reporting observational trials [13], and ARRIVE (Animal Research: Reporting of In Vivo Experiments) which provides a template for reporting animal studies [14].

Several of the approaches to grading or weighing evidence in the academic or grey literature also provide guidance on how to ensure clarity for the reader if evidence from a range of sources is being provided. For instance, Godfray et al. [15] outline the natural science evidence for bovine tuberculosis control with four labels highlighting the evidence source and strength[9].

In its methods for the development of public health guidance, NICE provides guidance on high-level evidence statements and templates of evidence tables for qualitative, quantitative and economic studies and reviews. These include quality and external validity scores (++, + or -), based on checklists also provided in the guidelines [7]. Such approaches can provide clarity in how conclusions have been formed when a specific body of literature is reviewed.

Interviewees stated that no prior information should be assumed when evidence is provided, and even basic details should be articulated. This includes providing a clear explanation of how a conclusion was reached and consideration of alternative hypotheses, to help guard against the conflation of correlation and causation.

6.2.2 Relevance and Reliability

EFSA defines reliability as how closely a piece of evidence represents the quantity, characteristic or event that it refers to, whereas the relevance of a piece of evidence can be defined as the contribution it would make to answering a specified question if it were completely reliable [3].

When judging the relevance of evidence, one can ask questions such as, how much extrapolation is required between the subjects and conditions that the evidence relates to, and those relevant for the question in hand? To answer this, understanding the context and the original question asked by those gathering that evidence is important [16]. This consideration was highlighted by interviewees as one of the most important factors for assessing whether evidence should be used in an assessment of food safety. The “appropriateness” of evidence to a policy question is emphasised by Parkhurst & Abeysinghe [17] and is one of DfID’s seven principles of high-quality research studies [8]. The Government’s Rapid Evidence Assessment Toolkit also highlights the importance of the relevance of the research design and study focus when assessing evidence [18].

International bodies such as the International Agency for Research on Cancer (IARC) and EFSA suggest key information that can help the assessment of relevance. This includes the context of the study; experimental conditions; details of

[9] [D_{ata}] A strong evidence base involving experimental studies or field data collection on bTB with appropriate detailed statistical or other quantitative analysis;
[E_{xp_op}] A consensus of expert opinion extrapolating results from other disease systems and well-established epidemiological principles;
[S_{upp_ev}] Some supporting evidence exists but further work would substantially improve the evidence base;
[P_{rojns}] Projections based on available evidence for which substantial uncertainty exists that could affect outcomes.

the substance being assessed; how confounding factors are dealt with; how the route of exposure was chosen; the size and nature of any effect, and how this was monitored. These are not only important questions in the planning stage of work [19], [20], but can also help in the assessment of whether the outcomes of a study are relevant to a particular policy question [20].

The language describing the original context and question should be precise. For instance, if a study was undertaken to assess whether the levels of a particular substance present in food are safe, this should include an explanation of how the safe levels were determined and the precision of the test, with reference to relevant national and international standards.

The use of international standards and practice, or methods widely used in the academic literature, can be a strong indicator of the reliability of evidence: the methodology is likely to have a known level of accuracy and the use of standard (validated) tests means that other researchers are likely to be able to repeat it. Acceptance of methods in international standards also indicates widespread expert opinion that a source of evidence is relevant for issues pertinent to food safety.

Any uncommissioned evidence that is submitted should refer to best practice including relevant guidance from national and international bodies, and methods widely referred to in the literature. Where an alternative approach is used, the reasons should be stated, and, where possible, the results compared with an accepted methodology.

Best practice guidance and methods for assessing key criteria of evidence exist for the different study types and components, including laboratory practice, studies on humans, modelling and data analysis.

Laboratory practice

Codex Alimentarius, the OECD, the WHO, and EFSA publish clear guidance and standards for laboratory practice; for example [21]–[26]. This includes the standards of Good Laboratory Practice (GLP) that laboratories must meet to facilitate the mutual acceptance of data in international trade. In the UK, testing facilities used to conduct studies on the safety of products for regulatory purposes must comply with GLP, assessed through the UK GLP compliance monitoring programme, led by the Medicines and Healthcare Products Regulatory Authority (MHRA).

While certification with standards such as GLP is not necessary for the acceptance of data outside regulatory processes, such standards and best practice outlined by national and international bodies provide a baseline for consideration of methods. For example, in its summary of how evidence is assembled and weighed, the IARC states that while “no critical evaluation or recommendation of any method is meant or implied”, emphasis is placed on those methods widely used for regulatory purposes [20].

The guidelines above include validated analytical tests, and quality control and assurance measures – which is important for ensuring that the outcomes of experiments are reliable [16], [27]. Standardised methodologies provide some

assurance about the reliability, accuracy and repeatability of the test. They also ensure that results can be compared to other similar agents, providing a greater understanding of the risk that they pose. This is important not only for ensuring that experiments are carried out well, but that they are relevant to the question being asked [27].

International standards for testing are developed by bodies such as the International Organization for Standardization (ISO) and European Committee for Standardization (CEN). Within the UK, accreditation for laboratory methods is provided by the UK Accreditation Service (UKAS).

In all instances the reason for choosing a particular method or study design must be provided; the use of an internationally accepted method is only fit for purpose when it is relevant to the issue being explored or the question being asked. For instance, the likely routes of exposure, magnitude of exposure, target organs and mode of action [19], [20], as well as the determination of biological sensitivity and specificity should be included to help assess biological relevance. In addition, the way in which an agent under assessment is prepared – or its form – is important for understanding if the tests being applied are appropriate. For example, nanomaterials can have very different biological interactions in comparison to their constituent chemicals and may require different tests for their characterisation and toxicological assessment [28].

One of the primary considerations in food safety is biological relevance. This is defined by EFSA as “an effect considered by expert judgement as important and meaningful for human, animal, plant or environmental health. It therefore implies a change that may alter how decisions for a specific problem are taken.” [19]. Providing the underlying data that fed into an expert judgement and outlining how widespread this opinion is and if there are any differing views within the relevant research community can indicate the strength of a judgement [29].

Assessment of biological relevance when using animal models is important, as the degree to which data from animal studies is relevant to humans, for instance when assessing a chemical hazard, will vary [30]. Results from animal experiments may be used alone or alongside epidemiological evidence of a risk to human health, and may be considered during a risk assessment where limited or no observational data on humans is available [20], [30]. Further, when assessing a risk to the health of a particular animal, evidence from other species may also be used. The use of appropriate animal models that will allow extrapolation, and an understanding of the uncertainties associated with this extrapolation is important for outlining the relevance of such a study [14], [19], [20]. In toxicology, the International Programme on Chemical Safety (IPCS) has published clear frameworks to aid the analysis of chemical modes of action in animal models and determine their relevance in humans for both cancer [31] and non-cancer [32] modes of action.

Studies on humans

Experimentation on, or observation of, human subjects can be helpful for gaining a full understanding of a food safety risk, where practically and ethically possible. When assessing the carcinogenic risk of specific chemicals to humans, the IARC has outlined the different types of study into cancer in humans that would be

considered in a review [20]. This outline provides guidance into the types of evidence that should be prioritised, and states that when other types of study are used a clear explanation of the rationale should be provided.

The guidelines for reporting studies in health research highlighted in the earlier section on Clarity also tend to include guidance on conduct, which will ensure that a study is reliably carried out [11], [12], [33]. EFSA recently published draft guidance on the appraisal of evidence from epidemiological studies, which includes experimental (or intervention) and non-experimental (observational) studies [34] and the FSA has its own guidelines on the synthesis of epidemiological studies developed by the Committees on Toxicity and Carcinogenicity [6]. This includes a checklist of components that should be present in all observational studies to be included in any meta-analysis or systematic review, adapted from MOOSE (Meta-analysis of Observational Studies in Epidemiology) [35].

As with laboratory studies, understanding the conditions and the context of the work can help in the assessment of whether an outcome is relevant – for example, providing details of the population studied, the measures used and what confounders were considered. NICE emphasise the importance of assessing external validity – that is, the extent to which findings from a study are generalisable to the whole source population [7]. The factors that may affect external validity can include variability within the population of the UK, for instance age, health and socio-economic status, but also the wider issue of whether the outcomes of a study conducted in one country are valid elsewhere. The target population (the population to which the study aims to draw inferences), the source population (the population used as the source of participants for the study), and the study population (the actual participants of the study) should be clearly articulated to assist the reader in assessing external validity [34].

Studies can also vary significantly with regards to the period over which they are conducted, so temporal effects should be clearly reported, as these can help the reader understand the relevance of a particular outcome, as well as providing detail on the potential mode(s) of action and exposure risk for any agent that has an impact on human health [3], [20].

Modelling

There are many instances where modelling may be required in addition to or alongside laboratory, clinical or observational data. EFSA [36], [37] and the Food and Drug Administration (FDA) [38] have published guidance on modelling that outlines approaches to the construction and use of models together with criteria for their reporting and evaluation. While the EFSA guidance is specifically focused on risk assessment for exposure to plant protection products and pesticide residues, the high-level criteria may be more broadly applied. A more detailed list of criteria for reporting modelling studies is outlined in Bennett & Manuel's review of reporting guidelines for modelling studies [39]. To understand if the model and its outputs are relevant, the problem definition, the model type and its structure are helpful to consider. This includes the underlying data being used to parameterise the model, the variables being tested, and the boundaries set, as well as the governing

equations being used and biological, chemical and physical properties that make up its structure.

EFSA's guidance on the use of probabilistic methodology for modelling dietary exposure to pesticide residues includes tables for choosing the type of model given the specific context of the assessment being performed [37]. WHO guidelines for chemical and pesticide exposure assessment provide widely used guidance for modelling dietary exposure using deterministic methods [40], [41]. As with laboratory and human studies, the use of widely recognised models recommended by national and international bodies can provide confidence in their relevance and reliability.

Further criteria that may be used to evaluate the reliability of a model include the plausibility of the science underpinning the model, the quality of the data underlying the model, and the correspondence of the model's outputs and behaviour during a test with independent observations [36]. These criteria can help to determine how closely the model resembles the system or process that it was designed to represent. The reliability of the outputs of the model should be tested using sensitivity analysis and all model outputs and their applications should include uncertainty analysis [3].

Spiegelhalter & Riesch provide an analysis of the different approaches that can be used to assess uncertainty in modelling [42]. Drawing on the experience of bodies such as the IPCC, they provide a five-level structure (Table 3) that outlines the different sources of uncertainties for within and between-model analysis and provide guidance on their expression.

Table 3 – types of uncertainty in models and their sources, reproduced from [42]

Level	What we are uncertain about	Source of uncertainty
1	Events	Essential unpredictability
2	Parameters within models	Limitations in information
3	Alternative model structures	Limitations in formalised knowledge
4	Effects of model inadequacy from recognised sources	Indeterminacy – known limitations in understanding and modelling ability
5	Effects of model inadequacy from unspecified sources	Ignorance – unknown limitations in understanding

While some of these uncertainties, such as the parameters within models, may be quantified, some are more appropriately expressed through a statement of confidence or an outline of the limitations of the model. Evidence derived from modelling should provide a comprehensive statement of uncertainty – including quantification where appropriate, and a statement of its limitations. Sensitivity testing can provide an indication of the between-model uncertainties arising from both parameters within models and alternative model structures.

Data analysis

As with data collection, the way that data is analysed affects its relevance to a particular question. Studies should include clear statements on how data was cleaned, processed and analysed, and why such approaches were taken. The raw data should be accessible whenever possible.

Appropriate statistical analysis is essential in scientific studies. Statistical significance may be used as an indicator of the strength of a particular piece of evidence but significance alone is insufficient for data to be important, meaningful or biologically relevant [27]. Statistical significance should only form one part of the statistical analysis of a well-designed experiment or study [19] and any particular statistical test must be used appropriately if it is to be meaningful. The use of significance tests is closely associated with hypothesis testing; the right power of study and significance test for one hypothesis is likely to be inappropriate for another [16], [19]. Care should be taken before using the outcome of statistical significance in one study to support an independent conclusion in another.

Any data showing a statistically significant effect should be accompanied by an explanation as to why the test used was appropriate as well as the biological relevance of the effect. EFSA recommend the use of statistical point estimation and confidence intervals alongside statistical significance, and state that P values should be quoted directly rather than simply stating a difference is “not significant” or , for example, “P<0.01”; EFSA also recommends access to the original data [19]. These approaches provide more information to the reader on the magnitude of any effect and its associated uncertainty.

Uncertainty and variability

An essential aspect of any data analysis is the uncertainty, which will help to indicate the reliability of an outcome [3]. All evidence must include consideration of uncertainty, including clear explanation of where uncertainties have come from, whether random variation, limitations of the apparatus, or epistemic uncertainty – that is, the effect of things that are simply unknown. The way these uncertainties are estimated or calculated, and their combination must be clearly explained [43].

The Codex Alimentarius Commission, FSA, WHO and EFSA provide clear guidelines on the calculation of uncertainties [43]–[48] and the approaches outlined in these documents should be followed wherever possible. Where appropriate, standard uncertainties should be quoted (e.g. [45]). When standard approaches are not used, an explanation of why should be provided. When using expert opinion, EFSA recommend an uncertainty level should be provided together with an explanation of the approach used for its estimate [49]. The Intergovernmental Panel on Climate Change (IPCC) produced a guidance note for the consistent treatment of uncertainties that is useful for the expression of uncertainty in the expert judgement of evidence compiled from various sources [50].

A related concept is that of variability. EFSA define variability as the “heterogeneity of values over time, space or different members of a population, including stochastic variability and controllable variability”. It tends to be differentiated from uncertainty as it may be reproducibly quantified [51] and cannot be reduced with greater

knowledge. Known sources should be explicitly outlined and may be quantified or controlled for [43]. Variability within a population can impact the generalisability of an assessment [51]; therefore, known variability should be taken into account when extrapolating from an experiment to a wider population.

6.3 Trust

For evidence to be trustworthy it must be honestly and openly communicated, so that it can be independently interrogated and any factors that could have influenced the results or the conclusions are disclosed. Two key criteria in ensuring this are transparency and impartiality.

6.3.1 Transparency

In all guidance assessed, transparency was a key factor in assessing the quality of evidence. It is important for the assessment of both reliability and relevance [52], and can help to reduce bias [53], [54]. In interviews, respondents described trustworthy evidence as being founded on transparency regarding what is known and unknown, why assumptions have been made (they will almost certainly have been required), and how gaps in knowledge are conveyed.

EFSA have published general principles used to guide the transparency of their own risk assessments, but which can be generally applied [52]. A clear outline of the source of the evidence, including the methods applied and the controls used, can help the reader understand the analysis applied and the rationale behind this. This aids assessment of the strength of the evidence and the relevance of both the tests and any conclusions proposed [3], [7], [20].

Where digital images, such as those produced through microscopy techniques or electrophoretic gels and blots, are used for analysis or to support a conclusion they should be treated as data. The means and conditions of capture should be clearly described in the methods together with any enhancement or processing techniques. The Office of Research Integrity in the United States provides guidelines on best practice in image processing [55].

An objective and impartial study should highlight any gaps in evidence, consider alternative hypotheses, and outline dissenting opinion or differing results from other publications or experts. If data have been omitted this should be stated and an explanation provided for why, so as to reduce cherry-picking. A clear statement on who did the work or provided the expert opinion and their funding sources can highlight potential sources of bias and therefore support its mitigation.

6.3.2 Impartiality and bias

Many guidelines recognise that the presence of bias and how it has been mitigated is fundamental in both the reliability of evidence and the reliability of its assessment. In the Grading of Recommendations, Assessment, Development and Evaluations

(GRADE) system for rating evidence, risk of bias is one of the key factors for which evidence is rated down; in the guidelines for using the system, key indicators of bias in study outcomes for both randomised controlled trials (RCTs) and observational studies are outlined [29]. Similarly, the IARC also considers the potential for bias in studies of cancer in humans to be key in understanding the quality of the study [20].

Mitigating partiality and bias was also highlighted as a key point during the interviews – both for those providing and those assessing evidence.

When assessing third-party evidence, interviewees highlighted the use of peer review within teams to mitigate individual bias. Interviewees highlighted instances where individuals receiving evidence have recognised bias in their judgement and sought to ensure it was mitigated. This has included passing work on to another member of the team for assessment where necessary.

Interviewees noted that the source of the evidence may be a factor that provides confidence in a study, but that the source or publisher of the evidence should not be the sole reason for its acceptance or dismissal. For instance, the importance of pre-prints during the Covid-19 pandemic was highlighted. Pre-prints should be subject to the same degree of critical appraisal as evidence that has been formally peer reviewed with any lack of formal review taken into account.

Some types of information fundamentally have increased risk of bias. For instance, there is inherent risk of bias in the use of expert judgement or opinion. However, this may be needed to support the interpretation of the available evidence – and may be included as a line of evidence in some evidence synthesis or weight of evidence approaches [15]. For instance, expert judgement may be required where there is insufficient data on a particular area in which a question is being asked or to which an outcome is being extrapolated [15], [49]. As highlighted in the section on relevance, expert judgement is necessary for assessing biological relevance [27], and is likely to be required for read-across from related materials when data on an agent is sparse.

In interviews, respondents characterised trustworthy expert judgement as ensuring that there is sufficient background evidence or data to allow the independent assessment of how the judgement was formed. Where there are gaps, this should be clearly stated and treated as an uncertainty.

Guidelines exist on the elicitation of expert judgement, including from EFSA [49] and other bodies outside of food safety, such as the IPCC [50], [56]. While care should be taken to ensure that the approach is proportionate, the underlying principles in this guidance provide a helpful reference. For instance, information can be provided on which experts were chosen, how and why; their funders and competing interests; what question was asked; whether there is any dissenting opinion; and what underlying data the judgement is based on. It should also be highlighted why expert opinion was sought rather than reliance on data [49]. This approach ensures transparency for the reader and can highlight any potential sources of bias, which can subsequently be mitigated against.

6.4 Robustness

Robustness encompasses the concepts of consistency and adequacy. The consistency of a result using the same and different methods provides confidence in its validity. A greater quantity of such evidence increases this confidence further. Adequacy is used in the context of the broad body of evidence that needs to be considered when assessing a risk or developing a policy or process. When new evidence is submitted on a subject it will be considered against this larger body of evidence.

6.4.1 Consistency

EFSA define consistency as the extent to which the contributions of different pieces of evidence to answering the specified question are compatible [3], while the US Agency for Healthcare Research and Quality (AHRQ) have described consistency as the extent to which similar findings are reported using similar and different study designs [57]. In their guidance on the use of the weight of evidence (WoE) approach in Risk Assessment, EFSA highlight that consistency is important when integrating different pieces of evidence, and is therefore important when comparing new pieces of evidence to the existing body of work [3].

This is not to say that a single study producing results that disagree with a larger body of studies that are consistent with each other should be disregarded. In the GRADE framework, inconsistency is listed as a factor that could lead to the downgrading of evidence quality rating, but a body of evidence is not up-rated if studies yield consistent results [58]. However, for a single well-defined question, only a single correct answer should be possible [3], therefore inconsistent evidence may suggest that at least one of the pieces of evidence is unreliable [8], [15], or that a factor important in the outcome has been missed. Such inconsistency will increase the uncertainty in a body of evidence, especially if the causes of the discrepancy are unknown.

Looking at the AHRQ definition, consistency can potentially be separated into two concepts: consistency of approach, which can be used to test the reproducibility of a piece of work, and consistency of outcome, given diverse methods of testing. Both concepts can provide confidence in the robustness of a result.

6.4.2 Reproducibility

Reproducibility is a core concept within science: if the outcome of an experiment is robust, it should be reproducible by other researchers under the same conditions as were first imposed [16]. Nevertheless, in 2016 a survey by Nature found that more than 70% of researchers have tried and failed to reproduce other scientists' work (with over half being unable to reproduce some of their own results) [54]. This undermines trust in the work and can weaken the field. For example, within cancer science, poor quality published pre-clinical data have previously been highlighted as a potential reason for failures in oncology trials [59]. Themes considered earlier in this review such as reliability, transparency and bias are important for bolstering reproducibility. Reproducible studies are likely to be those where researchers had taken care to describe the complete data set, appropriate controls and reliable

reagents were used, and investigator bias was taken into account, for example by asking investigators blinded to the experimental vs. control groups to perform the analysis [60]. If these actions were taken, they should be stated. These approaches may reduce cherry-picking of results, and reduce variation caused by differences in experimental conditions [60].

However, if the outcome of an experiment cannot be reproduced, it does not necessarily suggest an unreliable dataset, or incorrect conclusion – but may instead highlight an unknown variable that has not been controlled for [61]. In this way, repeating experiments under different conditions – with varied experimental conditions and methodologies – can either provide confidence in the robustness of a causal association, or help to provide the limits of an association and test alternative hypotheses.

6.4.3 Diversity of evidence

EFSA have highlighted that consistency intrinsically includes notions of both quantity and diversity of evidence, as the consistency of an outcome is accorded more weight when seen in a larger body of evidence, and when it is of diverse types [3]. Building up pieces of evidence from a range of sources that use different experimental methods, are applied to different populations, or alter different variables may provide a more complete picture of a food safety risk, and can be usefully integrated using methods such as meta-analysis for which there is extensive guidance [62], [63].

Using diverse methods and experimental practices to test the same hypothesis can also go some way to remove the limitations of the test itself. For example, a researcher can have more confidence that a result is unlikely to be due to an artefact of sample preparation or interaction with the testing equipment if the experiment is repeated using different methods [64]. Consistency of outcomes across different population groups, study designs and settings can provide confidence in generalisability of a result outside of the confines of a single study [3], [5].

Outside of the laboratory, large or complex systems may be tested with models, within which relationships will have been idealised, assumptions made about what should or should not be included, and different models for the same system may use alternative structures (see Table 1) [42]. This reflects both the limitations of researchers' knowledge, and a level of pragmatism – as all models are likely to be subject to limited time and resources [42], [65]. If different models, with varying assumptions and limitations, produce the same outcome when testing a hypothesis, it provides more confidence that this outcome is not sensitive to the limitations of the models tested [61].

6.4.4 Adequacy

The adequacy of a piece or body of evidence to influence a policy position or inform the evidence base is associated with both the quantity and strength of the evidence being assessed employing the criteria outlined in preceding sections of this paper. Clear guidelines exist on the adequacy of evidence within standardised risk assessment processes. For instance, the guidelines for regulated products outline the quantity and type of evidence that would be considered sufficient for a product to be brought onto the market (examples from EFSA [66] and the FSA [67]). Although

this review is about submissions made outside of these standardised processes, the content of existing guidelines can be helpful for assessing the adequacy of pieces of evidence.

When new evidence on an existing area is received FSA officials consider it in the context of the body of evidence that has already been used to develop a policy or inform a decision, as the FSA's food safety decisions are based upon a broad body of evidence. This is essential for determining whether any changes to the existing position may be needed.

In new, emerging or rapidly developing areas, a decision may need to be taken based upon limited evidence. In these circumstances, the FSA uses the best available evidence to make a decision, recognises where there are gaps or limitations in knowledge, and is open to change as new evidence becomes available.

Where evidence from a range of sources needs to be combined (including those of similar and different types), a WoE approach may be taken [3]. WoE is defined by the WHO as "a process in which all of the evidence considered to be relevant for a risk assessment is evaluated and weighted" [22]. WoE approaches facilitate the assessment of the adequacy of a set of evidence and can allow the comparison of different hypotheses – for instance, assessment based on a modified Bradford-Hill Framework can be used to compare different potential modes of action [68]. The frameworks used for weighing evidence may therefore be helpful when considering whether a piece of evidence is adequate for consideration, contributing to a body of evidence.

The WoE approach allows the combination of pieces of evidence that may otherwise be considered insufficient on their own, so that a judgement can be made considering the breadth of evidence and its combined weight. For example, while double-blind RCTs are the gold standard in clinical and population-based trials, there are instances when this approach cannot or should not be used. Another example may be that a study design is the first of its kind for a given question. This does not mean that the available evidence should be discounted; it might be used alongside other pieces of evidence to provide a body that is sufficient [7].

Often, evidence from a range of sources needs to be combined to provide a comprehensive assessment of an issue, as different types of evidence are relevant to different parts of the assessment. This may include different ways of gathering evidence within a field – such as quantitative and qualitative evidence, or evidence from different fields. For instance, toxicological and epidemiological evidence may need to be combined in order to produce a chemical risk assessment – an area in which the FSA's Committees on Toxicity and Carcinogenicity are currently developing guidelines [69]. During interviews it was highlighted that the FSA's integration of social sciences as a source of evidence in its risk analysis is a strength in this respect.

EFSA separate the process of a WoE assessment into three stages: assembling, weighing and integrating the evidence [3]. When uncommissioned evidence is submitted, assembling may be replaced with filtering; so the first step is to

understand whether the evidence is sufficient for consideration. The approaches used to weigh evidence, such as grading, can be used to aid filtering. Therefore, in this review only the assembling (filtering) and weighing steps will be considered, with the integration step coming into play only if a full risk analysis is commenced. It should be noted, however, that during a WoE assessment each piece of evidence will be considered individually during the assembling stage, whereas when uncommissioned evidence is received it is likely to be considered against the existing integrated body of evidence. Full reassessment of the evidence base for a policy or other food safety decision must be proportionate.

In its guidance on the use of a WoE assessment, EFSA divide the different types of assessment into three categories:

1. Best professional judgement – the qualitative integration of evidence using tools such as literature searches and non-quantitative systematic reviews;
2. Causal criteria – a criteria-based methodology for determining cause-effect relationships;
3. Rating – frameworks for rating evidence.

The approach to be used when filtering or weighing evidence depends significantly on the type of assessment to be performed; for instance, the use of Bradford Hill considerations (and its modifications) are based on causal criteria, and are widely recommended for epidemiological assessments [3], [70]. Any methodology chosen should be publicly available and ideally endorsed through wide usage or peer review [15].

Assessment approaches within causal criteria and rating may be helpful for filtering and weighing third-party evidence. One of the most widely used rating approaches is GRADE [71], [72], which rates evidence according to its certainty[10], outlined in Table 4.

Table 4 – GRADE certainty of evidence rating levels, reproduced from [71]

Certainty	What it means
Very low	The true effect is probably markedly different from the estimated effect
Low	The true effect might be markedly different from the estimated effect
Moderate	The authors believe that the true effect is probably close to the estimated effect
High	The authors have a lot of confidence that the true effect is similar to the estimated effect

[10] Initially the framework used the term “quality of evidence”, and somewhat different definitions for the grading levels were provided. [72]

The certainty of evidence can be downgraded due to risk of bias, imprecision, inconsistency, indirectness[11] and publication bias[12]. It should be noted that while GRADE provides a framework for rating evidence, like many other frameworks it remains subjective [71]. Additional frameworks and guidance have been published with regards to the above criteria [7], [74], [75] including detailed guidance by the GRADE working group [76], but each still require a level of expert judgement. In fact, some rating approaches have been criticised as being overly reliant on standardised approaches. The Occupational Health and Safety Administration (OSHA) of the USA highlights that care should be taken when using criteria developed by Klimisch et al. [77] for this reason [78].

To ensure that evidence is not arbitrarily filtered out because it does not use standardised methods, Beronius et al. developed a framework for rating non-standard in vivo studies, consisting of a two-tiered system for initially filtering and then weighing the evidence from a study based on its reliability and relevance [10]. This allows space for expert judgement, while ensuring that all the evidence underlying that judgement is clearly available.

In the social sciences, the INDEP (Index for Evidence in Policy) was produced for application to behavioural science [79], and has since been developed into the THEARI (Theoretical, Empirical, Applicable and Replicable Impact) rating system, which is designed to be more broadly applied (outlined in Table 5) [80]. Both of these systems rate the strength of evidence depending on its level of development, ranging from a scientifically viable concept or the identification of a potential issue, through to the translation of an intervention at scale with measurable impact.

Table 5 – THEARI validation levels for evidence, reproduced from [80]

Validation level	Rating	Description of Standard for Evidence
Theoretical	*	A scientifically viable concept has been proposed but lacks empirical testing or validation. May come in the form of a descriptive theory, explanation of an issue, or a framework of a wider construct. Opinions may be treated as theory.
Empirical	**	Insights exist that identify and explain a given issue using valid measurement of observation or phenomenon. Eventually it should include a move toward consensus on interpretations of robust study. May include non-successful interventions or lower-

[11] The four sources of indirectness being: differences in population, differences in interventions, differences in outcome measures and indirect comparisons [73]
[12] Defined by GRADE as a “a systematic under-estimation or … over-estimation of the underlying beneficial or harmful effect due to the selective publication of studies.” [73]

		power studies, with increasingly converging conclusions as new data are generated.
Applicable	***	Effective intervention or application completed, in a controlled trial where possible. Measurement of processes and effects considered valid. Effect should demonstrate value for scientific insight and/or practice via reasonably powered study. Ideally, the method was pre-registered for one or multiple studies.
Replicable	****	Valid and effective interventions produce converging conclusions through successful replication in terms of setting, procedure, and measurement. This is also a safeguard against errors (e.g. False positives) or bias tied to an individual study.
Impact	******	Successful translation of insight applied at scale, producing consistent and validated effects in line with prior conclusions. Findings validated at the highest conceivable power (ie. Populations) through real-world testing and replication of effects in multiple settings. Standard approach to implementation, evaluation and interpretation of data.

Considerations such as the magnitude of effect and the potential urgency of an intervention can influence both the weight of a given piece of evidence, and whether a body of evidence is considered sufficient for an intervention or policy decision to be made. For example, while in many instances both animal and human data would be required for an intervention to be made in humans, the requirement for additional evidence beyond animal studies may depend on the type of hazard identified. The IARC consider that in the absence of additional evidence, sufficient evidence of carcinogenicity in experimental animals means that it is biologically plausible that an agent would also present a risk to humans [20], and would lead the IARC to consider the agent as probably carcinogenic to humans. In addition, within GRADE, evidence may be upgraded given a large magnitude of effect or strong evidence of causality, such as a dose-response gradient [71].

The FSA's core remit is to ensure the safety of UK consumers. In interviews, respondents emphasised that the FSA may react more quickly or strongly to evidence that suggests a risk has been underestimated rather than overestimated. If an issue is urgent, there will be a trade-off between the speed of an assessment and its uncertainty. Communication between risk assessors and risk managers can help ensure that the approach is proportionate and appropriately thorough. Where evidence is insufficient, the FSA will use the precautionary principle to guide its approach, and preliminary conclusions will be reassessed as more evidence becomes available [81].

7 Conclusions

Best practice from the literature and from existing ways of working can be reviewed using the themes of quality, trust and robustness. Within these themes, detailed guidance and standards have been highlighted that can support officials and other experts' assessments of evidence in different fields, disciplines and contexts. Expert judgement and discretion remain important and are built into many of the frameworks reviewed here – from the assessment of biological relevance [19] to the recognition of the subjectivity of frameworks such as GRADE [71]. A single detailed checklist or grading system is considered to be inappropriate for the assessment of all the different types of evidence that the FSA sees. However, a set of overarching principles and guidelines based on the themes outlined here could be used to support initial decision-making and provide a shared understanding with third parties as to how their evidence can most effectively make a useful contribution to the FSA's body of evidence on a given topic, and provide an understanding of how their evidence will be evaluated.

8 Case Studies on the Provision and Consideration of Third-Party Evidence

The Science Council recognise that it may be hard to contextualise how the principles of quality, trust and robustness may in practice be applied to the appraisal of third-party evidence. It has sought to introduce two recent case studies that were highlighted during WP1 & 2 feedback and of relevance to topics raised during the WP4 public consultation.

8.1 Case Study One: Cannabidiol (CBD)

CBD is a phytocannabinoid chemical found in the *Cannabis* plant, belonging to the family of chemicals called cannabinoids. It has been researched for potential medical applications, including the treatment of epilepsy and seizures.

More recently, CBD has been used in non-medicinal products across a range of sectors including in food and drink and/or as a supplement. These products include beverages, edible oils, chewables and chocolates. In other sectors, topical creams and vaping liquids have been sold with CBD as an ingredient.

The formulation of these products can vary significantly. For instance, the broad range of CBD-containing foods and drinks available on the market, and their combination with different ingredients, is likely to affect their biological uptake. In addition, the precise composition of CBD products depends on the production and extraction methods used – meaning that some products could contain solvents or other cannabinoids such as tetrahydrocannabinol (THC) and cannabinol (CBN).

8.1.1 Relevance

As CBD food products started to emerge on the UK market, some of the best data that could be used to assess its safety was from trials of medical applications of CBD. From this evidence scientists in the FSA's Committee on Toxicity were able to assess the types of adverse effects on human health that CBD could cause, including hepatoxicity (liver injury), interactions with other medications and the lowest dose at which adverse effects had been observed.

However, the relevance of this data to food safety was limited by the difference in:

1. The original question being asked:
    - A balance between risks and benefits needs to be considered when assessing medical products. However, for food additives or novel foods, the assessment is only whether the product is safe – benefits do not need to be considered.
2. The formulation of the products:
    - The bioavailability of CBD is likely to vary significantly depending on whether it is taken with food and which type of food. The breadth of the market means that the ways in which CBD supplements and CBD-containing food products are prepared and their combination with other ingredients is wide ranging.

- Diverse extraction methods mean that other cannabinoids may be present in the products to a variable extent; this could cause interactions between the chemicals not considered in the evidence from medical trials.

3. How the product is consumed:
   - Medical CBD carries a recommended dosing regimen designed to balance the risks and benefits of the product, which should be monitored by health professionals. This is not the case for food products.

All producers using CBD in their products are now required to submit details via the regulated products process for novel foods. This ensures the evidence provided is relevant because it will need to include:

- The way in which the product was prepared and the composition of the final product – including other ingredients that may affect its biological uptake.
- Proposed uses for the specific product and anticipated intake, along with data on its absorption, distribution, metabolism and excretion given this context.
- Toxicological information and allergenicity, demonstrating that at the proposed levels of intake and given the proposed use, no significant adverse effects on human health are observed.

8.2 Case Study Two: Pathogenicity of STEC in raw milk cheese

In the summer of 2016, an outbreak of *E. coli* O157 occurred in Scotland that had a strong epidemiological link to the consumption of a raw milk cheese produced by a single business. Following the initial response to the outbreak, Local Authorities performed inspections to assess the food safety management system used during cheese production.

A variety of cheese types were sampled and tested for the presence of pathogenic bacteria. While the testing did not detect the bacterial strain implicated in the outbreak, it did detect several other non-O157 Shiga toxin-producing *E.coli* (STEC) in a number of cheeses produced by the business. Therefore, the Local Authority made the enforcement decision to withhold implicated products from the market.

As it was not possible to isolate the specific O157 outbreak strain in the cheeses, this decision was challenged. FSS was asked to provide evidence that the non-O157 STEC strains detected in the samples taken were pathogenic, i.e. capable of causing human illness.

8.2.1 Adequacy

In the first instance, evidence was sought to support the precautionary approach applied, while additional research was undertaken to understand the pathogenicity of the non-O157 STEC strains that had been detected.

The testing performed by the Local Authority had demonstrated that *E. coli* was being introduced into the cheese and that it was able to survive in the final product. In addition, epidemiologists at Public Health Scotland confirmed it was in fact very rare to find the strains isolated from clinical patients in implicated foodstuffs. This evidence supported the initial position taken by authorities, with risk management decisions based on epidemiological findings and food chain information.

FSS then undertook a series of tasks to investigate the pathogenicity (or otherwise) of the non-O157 STEC strains that had been detected.

Advice was sought from the Scottish *E. coli* O157/STEC Reference Laboratory (SERL) which specialises in the Whole Genome Sequencing of STEC in clinical infection in Scotland. It was confirmed that there had been a considerable number of cases of clinical non-O157 STEC infections in Scotland which had resulted in serious illness. Furthermore, they provided evidence that STEC containing similar pathogenic markers to those detected in the cheeses had been isolated from symptomatic patients.

8.2.2 Consistency

Consultation with experts at Public Health England led to the finding that strains with these pathogenic markers had also been identified in non-O157 STEC in human illness that had been reported in other parts of the UK.

A thorough literature review was conducted to collate evidence from research, sampling programmes and incidents on strains that had been identified in food and clinical isolates. The evidence base was developed further by consulting with Competent Authorities across Europe and the European STEC Reference Laboratory on their approach to assessing the pathogenicity of non-O157 STEC, and risk management actions taken when these STEC strains were identified in foods.

8.2.3 Outcome

This body of evidence was assimilated into a risk assessment which demonstrated that the strains isolated from the cheeses were theoretically capable of causing human illness. Local Authorities were able to use this when the business challenged the risk management decisions and actions taken.

9 Acknowledgements

The Science Council would like to thank the following people for their insight, input and commentary, which was essential for the preparation of this report:

- The FSA Chief Scientific Advisor, Professor Robin May
- FSA Officials in the Science, Evidence and Research Division (SERD) and Policy Directorate
- The FSA Science Advisory Committees, in particular the Synthesising Epidemiological and Toxicological Evidence (SETE) and Biological Relevance and Statistical Significance (BRSS) subgroups of the Committees on Toxicity, Carcinogenicity and Mutagenicity, and the Advisory Committee for Social Science
- Officials in other government departments who engaged, particularly those in the Chief Scientific Advisors' Offices.
- WP4 consultation respondents

10 Glossary

Term	Definition
Evidence	Any observational or experimental information used to support a hypothesis or position.
Quality	In this work we have defined quality using the concepts of clarity, relevance and reliability. This is evidence that is communicated effectively, can be applied to the question at hand, and where the outputs closely resemble the characteristic or event they refer to.
Clarity	Intelligible, coherent and cogent, with a logical thread running throughout.
Reliability	We use the EFSA definition of reliability: the extent to which the information comprising a piece or line of evidence is correct, i.e. how closely it represents the quantity, characteristic or event that it refers to.[3]
Relevance	We use the EFSA definition of relevance: the contribution a piece or line of evidence would make to answer a specified question, if the information comprising the line of evidence was fully reliable. In other words, how close is the quantity, characteristic or event that the evidence represents to the quantity, characteristic or event that is required in the assessment.[3]
Uncertainty	We use the definition of uncertainty provided by COT: as the estimated sum of the limits in knowledge [51], [82]. This includes limitations to apparatus, experimental techniques, models and study designs, as well as essential unpredictability [42].
Variability	Variability is defined as the inherent heterogeneity between individuals or groups, or over time or space. This may be humans, animals or other specimens. [3], [51]
Trust	The concept of trust is an expression of a combination of many of the other principles within this glossary, foremost of which is reliability. This is evidence where all the details of a study are openly and honestly communicated. It can be independently interrogated and any factors that could have influenced the results or conclusions are disclosed.
Transparency	Disclosure of all relevant information, including (but not limited to) data, gaps in data and omissions, and information regarding funding and interests. Openness to sharing data not originally disclosed when requested.
Impartiality	Recognition of the sources of bias and the action taken to mitigate them, including through methodological means and by disclosure, thereby ensuring that recipients of evidence are adequately aware of all factors that may have influenced the outcome of a study.
Robustness	In this work, robustness is defined using the concepts of consistency and adequacy.
Consistency	We use the AHRQ definition of consistency: the extent to which similar findings are reported using similar and different study

	designs [57], thereby encompassing the concepts of reproducibility and diversity of evidence.
Adequacy	The quantity and/or strength of evidence needed to support a policy position or inform an evidence base. This is dependent on considerations such as the existent evidence base, type of issue or question (i.e. Will this impact human health? How?) and magnitude of effect.
Reproducibility	The ability for a test or study to be carried out under the same or similar conditions and produce the same or a similar result.
Weight of evidence assessment	The process by which all evidence considered to be relevant to an assessment is evaluated (or weighted) and integrated to determine the relative support for possible answers to a scientific question. [3]

11 References

[1] ‘Government Analysis Function’, GOV.UK. https://www.gov.uk/government/organisations/government-analysis-function/about (accessed Jan. 19, 2021).

[2] ‘Our approach to science’, Food Standards Agency. https://www.food.gov.uk/about-us/our-approach-to-science (accessed Jan. 26, 2021).

[3] A. Hardy et al., ‘Guidance on the use of the weight of evidence approach in scientific assessments’, EFSA J., vol. 15, no. 8, p. e04971, 2017, doi: https://doi.org/10.2903/j.efsa.2017.4971.

[4] UK Statistics Authority, ‘Code of Practice for Statistics - About the Code’, Code of Practice for Statistics. https://code.statisticsauthority.gov.uk/the-code/ (accessed Jan. 23, 2021).

[5] Scientific Advisory Committee on Nutrition, ‘SACN: Framework for the Evaluation of Evidence’. Public Health England, 2006, Accessed: Dec. 04, 2020. [Online]. Available: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/872721/SACN_Framework_for_the_Evaluation_of_Evidence__March_2020_.pdf.

[6] Committee on Toxicity and Committee on Carcinogenicity, ‘Report of the Synthesising Epidemiological Evidence Subgroup (SEES) of the Committee on Toxicity and Committee on Carcinogenicity’, Food Standards Agency, 2018. Accessed: Jan. 19, 2021. [Online]. Available: https://webarchive.nationalarchives.gov.uk/20200808003610/https://cot.food.gov.uk/sites/default/files/seereportcotandcoc_0.pdf.

[7] The National Institute for Health and Care Excellence, ‘Methods for the development of NICE public health guidance (third edition)’. 2012, Accessed: Dec. 04, 2020. [Online]. Available: https://www.nice.org.uk/process/pmg4/resources/methods-for-the-development-of-nice-public-health-guidance-third-edition-pdf-2007967445701.

[8] The Department for International Development, ‘Assessing the Strength of Evidence’, The Department for International Development, 2014. [Online]. Available: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/291982/HTN-strength-evidence-march2014.pdf.

[9] ‘Guidance on Statistical Reporting’, EFSA J., vol. 12, no. 12, p. 3908, 2014, doi: https://doi.org/10.2903/j.efsa.2014.3908.

[10] A. Beronius, L. Molander, C. Rudén, and A. Hanberg, ‘Facilitating the use of non-standard in vivo studies in health risk assessment of chemicals: a proposal to improve evaluation criteria and reporting’, J. Appl. Toxicol., vol. 34, no. 6, pp. 607–617, 2014, doi: https://doi.org/10.1002/jat.2991.

[11] EQUATOR, ‘The EQUATOR Network | Enhancing the QUAlity and Transparency Of Health Research’, The EQUATOR Network. https://www.equator-network.org/ (accessed Dec. 04, 2020).

[12] CONSORT, ‘CONsolidated Standards of Reporting Trials Guideline’, CONSORT-2010, 2010. http://www.consort-statement.org/consort-2010.

[13] E. von Elm, D. G. Altman, M. Egger, S. J. Pocock, P. C. Gøtzsche, and J. P. Vandenbroucke, ‘Strengthening the reporting of observational studies in

epidemiology (STROBE) statement: guidelines for reporting observational studies', BMJ, vol. 335, no. 7624, pp. 806–808, Oct. 2007, doi: 10.1136/bmj.39335.541782.AD.
[14] N. Percie du Sert et al., 'Reporting animal research: Explanation and elaboration for the ARRIVE guidelines 2.0', PLOS Biol., vol. 18, no. 7, p. e3000411, Jul. 2020, doi: 10.1371/journal.pbio.3000411.
[15] H. C. J. Godfray et al., 'A restatement of the natural science evidence base relevant to the control of bovine tuberculosis in Great Britain †', Proc. R. Soc. B Biol. Sci., vol. 280, no. 1768, p. 20131634, Oct. 2013, doi: 10.1098/rspb.2013.1634.
[16] C. D. Collins, M. Baddeley, G. Clare, R. Murphy, S. Owens, and S. Rocks, 'Considering evidence: The approach taken by the Hazardous Substances Advisory Committee in the UK', Environ. Int., vol. 92–93, pp. 565–568, Jul. 2016, doi: 10.1016/j.envint.2016.01.006.
[17] S. Abeysinghe and J. Parkhurst, '"Good" evidence for improved policy making: from hierarchies to appropriateness', Jan. 2013.
[18] Government Social Research Service, 'How to do a Rapid Evidence Assessment [ARCHIVED CONTENT]'. https://webarchive.nationalarchives.gov.uk/20140402163101/http://www.civilservice.gov.uk/networks/gsr/resources-and-guidance/rapid-evidence-assessment/how-to-do-a-rea (accessed Jan. 23, 2021).
[19] B. Antunovic, 'Statistical Significance and Biological Relevance', EFSA J., vol. 9, no. 9, p. 2372, 2011, doi: https://doi.org/10.2903/j.efsa.2011.2372.
[20] IARC, 'IARC Monographs on the Evaluation of Carcinogenic Risks to Humans - Preamble'. World Health Organization International Agency for Research on Cancer, 2006, [Online]. Available: https://monographs.iarc.fr/wp-content/uploads/2018/06/CurrentPreamble.pdf.
[21] M. Thompson and R. Wood, 'Harmonized guidelines for internal quality control in analytical chemistry laboratories (Technical Report)', Pure Appl. Chem., vol. 67, no. 4, pp. 649–666, Jan. 1995, doi: 10.1351/pac199567040649.
[22] Principles and methods for the risk assessment of chemicals in food. World Health Organization, 2008.
[23] Codex Alimentarius, 'GUIDELINES ON GOOD LABORATORY PRACTICE IN PESTICIDE RESIDUE ANALYSIS (CAC/GL 40-1993)'. Revision . Amendment 2010 2003, Accessed: Dec. 04, 2020. [Online]. Available: http://www.fao.org/fao-who-codexalimentarius/sh-proxy/en/?lnk=1&url=https%253A%252F%252Fworkspace.fao.org%252Fsites%252Fcodex%252FStandards%252FCXG%2B40-1993%252Fcxg_040e.pdf.
[24] Environment Directorate, 'OECD Principles on Good Laboratory Practice'. OECD, Accessed: Dec. 04, 2020. [Online]. Available: http://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?cote=env/mc/chem(98)17&doclanguage=en.
[25] Joint FAO/WHO Expert Meetings on Microbiological Risk Assessment (JEMRA) on Methodologies of Microbiological Risk Assessment, 'Draft Guidance of Microbiological Risk Assessment for Food', World Health Organization; Food and Agriculture Organizations of the United Nations, 2020. Accessed: Jan. 19, 2021. [Online]. Available: https://www.who.int/docs/default-source/food-safety/jemra/call-for-consultation/methodology-report-public-comments.pdf?sfvrsn=bf30c148_2.

[26] World Health Organization and Food and Agriculture Organization of the United Nations, Eds., Hazard characterization for pathogens in food and water: guidelines. Geneva, Switzerland : Rome: World Health Organization ; Food and Agriculture Organization of the United Nations, 2003.
[27] A. Hardy et al., 'Guidance on the assessment of the biological relevance of data in scientific assessments', EFSA J., vol. 15, no. 8, p. e04970, 2017, doi: https://doi.org/10.2903/j.efsa.2017.4970.
[28] A. Hardy et al., 'Guidance on risk assessment of the application of nanoscience and nanotechnologies in the food and feed chain: Part 1, human and animal health', EFSA J., vol. 16, no. 7, p. e05327, 2018, doi: https://doi.org/10.2903/j.efsa.2018.5327.
[29] G. H. Guyatt et al., 'GRADE guidelines: 4. Rating the quality of evidence--study limitations (risk of bias)', J. Clin. Epidemiol., vol. 64, no. 4, pp. 407–415, Apr. 2011, doi: 10.1016/j.jclinepi.2010.07.017.
[30] E. Aiassa, 'Principles and process for dealing with data and evidence in scientific assessments', EFSA J., vol. 13, no. 6, p. 4121, 2015, doi: https://doi.org/10.2903/j.efsa.2015.4121.
[31] A. R. Boobis et al., 'IPCS Framework for Analyzing the Relevance of a Cancer Mode of Action for Humans', Crit. Rev. Toxicol., vol. 36, no. 10, pp. 781–792, Jan. 2006, doi: 10.1080/10408440600977677.
[32] A. R. Boobis et al., 'IPCS Framework for Analyzing the Relevance of a Noncancer Mode of Action for Humans', Crit. Rev. Toxicol., vol. 38, no. 2, pp. 87–96, Jan. 2008, doi: 10.1080/10408440701749421.
[33] European Medicines Agency, 'E 3 Structure and Content of Clinical Study Reports', p. 48, 2006.
[34] S. More et al., 'Draft for internal testing Scientific Committee guidance on appraising and integrating evidence from epidemiological studies for use in EFSA's scientific assessments', EFSA J., vol. 18, no. 8, p. e06221, 2020, doi: https://doi.org/10.2903/j.efsa.2020.6221.
[35] D. F. Stroup et al., 'Meta-analysis of observational studies in epidemiology: a proposal for reporting. Meta-analysis Of Observational Studies in Epidemiology (MOOSE) group', JAMA, vol. 283, no. 15, pp. 2008–2012, Apr. 2000, doi: 10.1001/jama.283.15.2008.
[36] A. Aagaard, 'Scientific Opinion on good modelling practice in the context of mechanistic effect models for risk assessment of plant protection products', EFSA J., vol. 12, no. 3, p. 3589, 2014, doi: https://doi.org/10.2903/j.efsa.2014.3589.
[37] J. Boeston, 'Guidance on the Use of Probabilistic Methodology for Modelling Dietary Exposure to Pesticide Residues', EFSA J., no. 2012;10(10):2839, doi: 10.2903/j.efsa.2012.2839.
[38] 'Reporting of Computational Modeling Studies in Medical Device Submissions - Guidance for Industry and Food and Drug Administration Staff', p. 48.
[39] C. Bennett and D. G. Manuel, 'Reporting guidelines for modelling studies', BMC Med. Res. Methodol., vol. 12, no. 1, p. 168, Nov. 2012, doi: 10.1186/1471-2288-12-168.
[40] Global Environment Monitoring System and Codex Alimentarius, 'Guidelines for predicting dietary intake of pesticide residues', World Health Organization, 1997. Accessed: Jan. 21, 2021. [Online]. Available: https://www.who.int/foodsafety/publications/chem/en/pesticide_en.pdf?ua=1.

[41] S. Joint FAO/WHO Consultation on Food Consumption and Exposure Assessment of Chemicals (1997: Geneva, W. H. O. P. of F. S. and F. Aid, and F. and A. O. of the U. Nations, 'Food consumption and exposure assessment of chemicals : report of a FAO/WHO consultation, Geneva, Switzerland, 10-14 February 1997', 1997, Accessed: Jan. 21, 2021. [Online]. Available: https://apps.who.int/iris/handle/10665/63988.

[42] D. J. Spiegelhalter and H. Riesch, 'Don't know, can't know: embracing deeper uncertainties when analysing risks', Philos. Trans. R. Soc. Math. Phys. Eng. Sci., vol. 369, no. 1956, pp. 4730–4750, Dec. 2011, doi: 10.1098/rsta.2011.0163.

[43] D. Benford et al., 'Guidance on Uncertainty Analysis in Scientific Assessments', EFSA J., vol. 16, no. 1, p. e05123, 2018, doi: https://doi.org/10.2903/j.efsa.2018.5123.

[44] Codex Alimentarius, 'GUIDELINES ON MEASUREMENT UNCERTAINTY (CXG 54-2004)'. 2004, [Online]. Available: http://www.fao.org/fao-who-codexalimentarius/sh-proxy/en/?lnk=1&url=https%253A%252F%252Fworkspace.fao.org%252Fsites%252Fcodex%252FStandards%252FCXG%2B54-2004%252FCXG_054e.pdf.

[45] Codex Alimentarius, 'GUIDELINES ON ESTIMATION OF UNCERTAINTY OF RESULTS (CAC/GL 59-2006)'. Amended 2011 2006, Accessed: Dec. 04, 2020. [Online]. Available: http://www.fao.org/fao-who-codexalimentarius/sh-proxy/en/?lnk=1&url=https%253A%252F%252Fworkspace.fao.org%252Fsites%252Fcodex%252FStandards%252FCXG%2B59-2006%252Fcxg_059e.pdf.

[46] Andy Hart, John Paul Gosling, Alan Boobis, David Coggon, Peter Craig, and David Jones, 'Development of a framework for evaluation and expression of uncertainties in hazard and risk assessment', The Food and Environment Research Agency, 2010. Accessed: Jan. 22, 2021. [Online]. Available: https://www.food.gov.uk/research/research-projects/development-of-a-framework-for-evaluation-and-expression-of-uncertainties-in-hazard-and-risk-assessment.

[47] World Health Organization, Guidance document on evaluating and expressing uncertainty in hazard characterization. World Health Organization, 2018.

[48] World Health Organization, Uncertainty and data quality in exposure assessment. World Health Organization, 2008.

[49] Bolger, Fergus, 'Guidance on Expert Knowledge Elicitation in Food and Feed Safety Risk Assessment', EFSA J., vol. 12, no. 6, p. 3734, 2014, doi: https://doi.org/10.2903/j.efsa.2014.3734.

[50] Intergovernmental Panel on Climate Change, 'IPCC Cross-Working Group Meeting on Consistent Treatment of Uncertainties', Intergovernmental Panel on Climate Change, 2010. Accessed: Jan. 22, 2021. [Online]. Available: https://www.ipcc.ch/publication/ipcc-cross-working-group-meeting-on-consistent-treatment-of-uncertainties/.

[51] Committee on Toxicity, 'COT report on variability and uncertainty in toxicology', Food Standards Agency, 2007. Accessed: Jan. 22, 2021. [Online]. Available: https://webarchive.nationalarchives.gov.uk/20200803163700/https://cot.food.gov.uk/cotreports/cotwgreports/cotwgvut.

[52] Barlow, Susan, 'Guidance of the Scientific Committee on Transparency in the Scientific Aspects of Risk Assessments carried out by EFSA. Part 2: General Principles', EFSA J., vol. 7, no. 5, p. 1051, 2009, doi: https://doi.org/10.2903/j.efsa.2009.1051.

[53] R. Nuzzo, ‘How scientists fool themselves - and how they can stop’, Nature, vol. 526, no. 7572, pp. 182–185, Oct. 2015, doi: 10.1038/526182a.
[54] M. Baker, ‘1,500 scientists lift the lid on reproducibility’, Nature, vol. 533, no. 7604, pp. 452–454, 26 2016, doi: 10.1038/533452a.
[55] ‘Guidelines for Best Practices in Image Processing’. https://ori.hhs.gov/education/products/RIandImages/guidelines/list.html (accessed Jan. 22, 2021).
[56] IPCC, ‘Chapter 2: Approaches to Data Collection’, in 2006 IPCC Guidelines for National Greenhouse Gas Inventories: General Guidance and Reporting, vol. 1, IPCC, 2006, pp. 2.1-2.24.
[57] S. West et al., Systems to Rate the Strength of Scientific Evidence: Summary. Agency for Healthcare Research and Quality (US), 2002.
[58] G. H. Guyatt et al., ‘GRADE guidelines: 7. Rating the quality of evidence—inconsistency’, J. Clin. Epidemiol., vol. 64, no. 12, pp. 1294–1302, Dec. 2011, doi: 10.1016/j.jclinepi.2011.03.017.
[59] C. G. Begley and L. M. Ellis, ‘Raise standards for preclinical cancer research’, Nature, vol. 483, no. 7391, Art. no. 7391, Mar. 2012, doi: 10.1038/483531a.
[60] M. Baker, ‘How quality control could save your science’, Nature, vol. 529, no. 7587, pp. 456–458, Jan. 2016, doi: 10.1038/529456a.
[61] K. J. Rothman and S. Greenland, ‘Hill’s Criteria for Causality’, in Encyclopedia of Biostatistics, American Cancer Society, 2005.
[62] Centre for Reviews and Dissemination, Systematic Reviews: CRD’s guidance for undertaking reviews in health care. 2009.
[63] Higgins JPT et al., Cochrane Handbook for Systematic Reviews of Interventions, 6.1. Cochrane, 2020.
[64] J. N. Schupbach, ‘Robustness Analysis as Explanatory Reasoning’, Br. J. Philos. Sci., vol. 69, no. 1, pp. 275–300, Mar. 2018, doi: 10.1093/bjps/axw008.
[65] A. Hardy et al., ‘Editorial: Increasing robustness, transparency and openness of scientific assessments’, EFSA J., vol. 13, no. 3, p. e13031, 2015, doi: https://doi.org/10.2903/j.efsa.2015.e13031.
[66] ‘Data requirements for the evaluation of food additive applications’, EFSA J., vol. 7, no. 8, p. 1188, 2009, doi: https://doi.org/10.2903/j.efsa.2009.1188.
[67] ‘The UK transition’, Food Standards Agency. http://food.gov.uk/business-guidance/the-uk-transition (accessed Dec. 04, 2020).
[68] M. E. (Bette) Meek, C. M. Palermo, A. N. Bachman, C. M. North, and R. Jeffrey Lewis, ‘Mode of action human relevance (species concordance) framework: Evolution of the Bradford Hill considerations and comparative analysis of weight of evidence’, J. Appl. Toxicol., vol. 34, no. 6, pp. 595–606, Jun. 2014, doi: 10.1002/jat.2984.
[69] ‘SETE | Committee on Toxicity’. https://cot.food.gov.uk/SETEworkinggroup (accessed Jan. 19, 2021).
[70] K. N. Lohr, ‘Rating the strength of scientific evidence: relevance for quality improvement programs’, Int. J. Qual. Health Care, vol. 16, no. 1, pp. 9–18, Feb. 2004, doi: 10.1093/intqhc/mzh005.
[71] ‘What is GRADE? | BMJ Best Practice’. https://bestpractice.bmj.com/info/toolkit/learn-ebm/what-is-grade/ (accessed Dec. 04, 2020).
[72] M. Hultcrantz et al., ‘The GRADE Working Group clarifies the construct of certainty of evidence’, J. Clin. Epidemiol., vol. 87, pp. 4–13, Jul. 2017, doi: 10.1016/j.jclinepi.2017.05.006.

[73] Holger Schunemann, Jan Brozek, Gordon Guyatt, and Andrew Oxman, GRADE handbook. 2013.
[74] J. P. T. Higgins et al., 'The Cochrane Collaboration's tool for assessing risk of bias in randomised trials', BMJ, vol. 343, Oct. 2011, doi: 10.1136/bmj.d5928.
[75] J. A. Sterne et al., 'ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions', BMJ, vol. 355, Oct. 2016, doi: 10.1136/bmj.i4919.
[76] GRADE Working Group, 'Journal of Clinical Epidemiology GRADE series', Journal of Clinical Epidemiology. https://www.jclinepi.com/content/jce-GRADE-Series (accessed Jan. 18, 2021).
[77] H. J. Klimisch, M. Andreae, and U. Tillmann, 'A systematic approach for evaluating the quality of experimental toxicological and ecotoxicological data', Regul. Toxicol. Pharmacol. RTP, vol. 25, no. 1, pp. 1–5, Feb. 1997, doi: 10.1006/rtph.1996.1076.
[78] Occupational Safety and Health Administration, 'Guidance on Data Evaluation for Weight of Evidence Determination: Application to the 2012 Hazard Communication Standard'. United States Department of Labor, 2012, [Online]. Available: https://www.osha.gov/weightofevidence/woe_questions.html.
[79] K. Ruggeri et al., Behavioral Insights for Public Policy: Concepts and Cases, 1st ed. Routledge, 2018.
[80] Kai Ruggeri et al., 'Standards for evidence in policy decision-making', Nat. Res. Soc. Behav. Sci., 2020, Accessed: Jan. 22, 2021. [Online]. Available: go.nature.com/2zdTQIs.
[81] Guy Poppy, 'Chief Scientific Advisor's Report on Risk Analysis', Food Standards Agency, 2020. Accessed: Jan. 19, 2021. [Online]. Available: https://www.food.gov.uk/sites/default/files/media/document/csa-report-risk-analysis.pdf#page=14.
[82] Science Council, 'Report of the Science Council's Working Group on risk and uncertainty', Food Standards Agency, 2018. Accessed: Feb. 16, 2021. [Online]. Available: https://webarchive.nationalarchives.gov.uk/20200803142331/https://science-council.food.gov.uk/science-council-subgroups/science-council-working-group-on-risk-and-uncertainty.