CC/MUT/2020/03

COMMITTEES ON CARCINOGENICITY/MUTAGENICITY/TOXICITY OF CHEMICALS IN FOOD, CONSUMER PRODUCTS AND THE ENVIRONMENT (COC/COM/COT)

SCOPING DOCUMENT: BIOLOGICAL RELEVANCE AND STATISTICAL SIGNIFICANCE

Background

1. The topic of 'biological relevance and statistical significance' has been raised as an area of interest during Committee horizon scanning activities. Following discussion at the January 2020 COT meeting, it was decided that a scoping paper would be prepared for joint consideration by the COT, COC, and COM.

2. The horizon scanning item presented at the January 2020 COT meeting (TOX/2020/09) defined the issue as follows: "In terms of priorities for joint Committee consideration, it was suggested one important area was how to evaluate the biological or toxicological relevance of a reported response or perturbation, especially where this may be an atypical endpoint and how statistics can, and should, be used to help determine this… …This should encompass how the Committees could judge whether the statistics used were appropriate. Consideration of sufficient levels of health protection and dealing with uncertainty could also be useful, for example, the degree of confidence over a non-significant result in relation to health protection.".

Introduction

3. The problem of determining how experimental results can best be judged to establish their importance, or rather the chance of them being 'significant' in relation to the issue being investigated, has led to a somewhat philosophical but important debate relating to the relationship between biological relevance and statistical significance.

4. The judgment of statistical significance is typically based on an assessment of the likelihood[1] of the observed effect occurring by chance alone; it is normally assumed that any result with a less than 5% probability of occurring by chance (P<0.05) is the consequence of a causal relationship between intervention (experimental variable) and outcome (observation), rather than random variation.

[1] Note that the word 'likelihood' is used here in its general meaning in English, not the specific statistical sense where it is defined as "the probability of a set of observations given the value of some parameter or set of parameters".

This choice of statistical 'threshold' is of course subjective and may vary. The other side of the equation is biological relevance. A result that is statistically significant may not be judged to be biologically relevant, or one that does not reach statistical significance may be of potential biological relevance; but how is this judgement of biological relevance best made? How can questions of subjectivity be avoided or mitigated?

5. It may be asserted that a result that is biologically relevant but lacks statistical significance is not important; in other words, it is not possible to determine whether there is a response in the absence of some consideration of chance variation. But this raises questions of level of protection – how confident do we need to be that an effect would be below a certain incidence or magnitude. Equally it can be argued that a statistically significant finding is valueless if it is not biologically relevant.

6. A conflict may occur where an effect is judged to have extreme biological relevance (and importance), e.g. teratogenicity, which is in line with a known or suspected Mode of Action (MoA), for example, but a standard statistical analysis (using $P<0.05$) fails, for whatever reason, to demonstrate a statistically significant difference. In this case there would be a strong driver to investigate thoroughly the statistical result and perhaps adjust the chosen threshold of significance, or argue that the finding of biological relevance should be given a stronger weighting than the statistics, or ensure that any advice provided reflects the uncertainty and its implications for possible outcomes. This highlights the inherent lack of 'certainty' in such assessments and the need to fully consider both elements.

7. Determining a robust approach to the assessment and interpretation of biological relevance and statistical significance is – or should be – fundamental to the work of all experimental scientists and applies especially to expert committees and other bodies tasked with making critical judgments on, for example, consequences to public health of exposure to noxious substances.

8. This scoping paper summarises some of the more relevant and significant work that has been published on this issue in recent years, including some of the methodologies and definitions that have been developed. The document is based primarily on four literature sources: two guidance documents published by the European Food Safety Authority (EFSA) and two journal publications by Lovell. The four sources are addressed in turn, and thus there is some unavoidable repetition in the aspects considered. Additional literature sources of relevance are in general not cited but can be found within the four publications on which this present document is based. Some general comments on the themes considered in this scoping paper, based on comments received from the COT, COM and COC Chairs as part of the preparation of this paper is presented at the end of the paper (paragraphs 54 – 58).

EFSA (2011) Scientific Opinion on Statistical Significance and Biological Relevance[2]

9. In 2011, the EFSA Scientific Committee published a scientific opinion on 'Statistical Significance and Biological Relevance' (EFSA 2011). This work was developed by an EFSA Working Group, with the objective to help EFSA Scientific Panels and Committee in the assessment of biologically relevant effects.

10. EFSA noted that although there existed substantial expertise in toxicological hazard identification and risk assessment relating to chemicals, including standardised methods and guidelines for the conduct of toxicology studies (e.g. Organisation for Economic Co-operation and Development (OECD) guidelines), there was still some debate about methods for conducting, analysing, and interpreting such studies. The panel noted that risk assessment of individual chemicals could require specific studies relating to mechanism or mode of action, and there was a need to address carefully the area of design and statistical interpretation of such studies. For statistical analysis, it was noted in particular a possibility for over-reliance on the use of specific probability levels to indicate either positive or negative effects, something that had been commented on previously by statisticians. The panel noted that approaches used in the pharmaceutical industry might be usefully applied to toxicology studies; for example, concepts of bioequivalence/inferiority/superiority testing.

11. The EFSA panel explored the concepts of 'biological relevance' and 'statistical significance' and the relationship between the two.

12. The interpretation of study data may be limited by a lack of standards to define quantitative changes which designate biological relevance, and this can impact decisions as to whether effects observed in studies, for example histopathological changes in toxicological studies, are considered to be biologically important. The following meaning was proposed for biological relevance:

- "*A biologically relevant effect can be defined 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*."

13. This description would assume the existence of a 'normal' biological state, and judgement of what would be considered as a biologically relevant effect should be made by experts in the particular field of investigation. Ideally, consideration of the size of effect that would be taken as biologically relevant should be made at the design stage of a study, although in reality this may not always be practical.

14. The Committee went on to define statistical significance as follows:

[2] https://www.efsa.europa.eu/en/efsajournal/pub/2372 (accessed 15/07/2020)

- "*Statistical significance is a measure of how likely an observed result could have occurred, on the basis of a set of assumptions. (Reese, 2004).*"

15. Statistical testing often leads to categorisation of findings as either 'significant' or 'non-significant'. These concepts derive from the framework of statistical hypothesis testing, most commonly underlaid by a combination of the Neyman-Pearson and Fisherian paradigms[3] (see paragraph 37). Two hypotheses are tested: the null hypothesis and an alternative hypothesis and, based on the outcome of tests applied, are either rejected or not rejected. A type I error (α) occurs if a true null hypothesis is rejected. A type II error (β) occurs if a false null hypothesis is accepted. Effects that are found are reported as *P*-values, being the probability that an effect of at least the magnitude observed would have occurred by chance alone when the null hypothesis is in fact true. As the type I error α indicates the probability of rejecting a null hypothesis that is true, this value can be used as the cut-off point for significance testing (the threshold of significance, or 'critical value'); this should be chosen in advance of performing the test, ideally during study design. Tests that produce a *P*-value lower than the pre-determined critical value are often termed 'significant', based on the (low) level of probability of observing such an outcome if the null hypothesis is true. Statistical power is the probability of identifying a pre-defined effect if it actually exists (correctly rejecting the null hypothesis; 1-β) and is often set at 80% (β=0.2). Power is dependent on the study sample size. The EFSA Scientific Committee noted that power analyses are sometimes made retrospectively after a study has been carried out, but the Committee concluded that this practice is not acceptable and should not be recommended.

16. In exploring the relationship between biological relevance and statistical significance, the EFSA panel noted in particular the confusion that can occur with the use of the term 'significant'. While this term is often used in general language to indicate large-size or relevance, this is not the case in the statistical meaning, but nevertheless the term 'statistically significant' in reporting the analysis of study data is sometimes incorrectly taken to imply effect size or biological relevance. Furthermore, a finding of statistical significance may be assumed to represent mathematical 'proof' of a biologically relevant effect, but the EFSA panel considered that establishment of biological relevance should be the primary factor in the assessment and not the specific level of statistical significance.

17. Further discussion of interpretation of statistical significance focusses on potential errors. The EFSA panel make the point that 'absence of evidence is not evidence of absence'. Breaking down this statement, 'absence of evidence' refers to non-rejection of a null hypothesis specifying no given effect, usually with $P > 0.05$ or $P > 0.01$; i.e. there is no evidence of an effect. Conversely, 'evidence of absence' relates to a null hypothesis of effect, whereby a statistically significant outcome can indicate rejection of the null hypothesis; i.e. evidence of no effect. The requirement

[3] This methodology is commonly referred to as 'null hypothesis statistical testing' (NHST) and is described further in the later section, 'Statistical approaches and their limitations'.

for 'evidence of absence' requires careful consideration, and this concept has been addressed in the pharmaceutical sector in 'equivalence testing' of generic drugs: here the requirement is placed on the demonstration of equivalence by finding significant evidence against a null hypothesis of non-equivalence (no different in effect between generic and parent drug). These concepts were highlighted by EFSA in a table that is reproduced below (Table 1):

Table 1. Summary table on absence of evidence and evidence of absence. The two concepts apply in two different contexts: *difference tests* and *equivalence tests*, respectively. The column on the right (Outcome) stresses the fact that a *P*-value above the set cut-off does not allow any conclusion to be drawn. Reproduced from Table 2 of EFSA (2011).

	H_0: no effect (difference test)	**H_0: effect (equivalence test)**	**Outcome**
***P* < the chosen threshold of significance**	Evidence of presence	Evidence of absence	A conclusion can be drawn
***P* > the chosen threshold of significance**	Absence of evidence	Absence of evidence	No conclusion can be drawn

18. An important point is the consideration of statistical analysis (as opposed to statistical significance). Statistical analysis is conducted to explore possible relationships, patterns, trends and/or make inferences, while statistical significance simply reports the outputs of tests conducted (usually *P*-values). As mentioned in paragraph 6, the EFSA Committee considered biological relevance to be the primary factor of importance. The Committee noted that even if a statistically significant result is obtained, the biological effect size may be too small to be relevant, and as such all data from statistical analyses (i.e. not just statistical significance) should also be considered in the context of biological relevance.

19. The EFSA Committee considered that the calculation of a biological effect using a statistical point estimate and its uncertainty (interval estimate; confidence intervals (CIs)) provides more information than the simple result of a significance test. The use of CIs reflects uncertainty in the dataset, and wide CIs indicate a lack of information (e.g. small sample size). The Committee considered that use of CIs helps to avoid an absolute cut-off between 'yes/no' at, for example, *P* >0.05, and can be informative in conjunction with nonsignificant results.

20. The EFSA Scientific Committee made the following recommendations in its report on biological relevance and statistical significance:

- "The distinction between *biological relevance* and *statistical significance* should be acknowledged when developing scientific opinions

- "Where possible, the relevant biological effect and its desired size should be considered at an early stage of study design and the plan for assessment

- "The term *significance/significant* should be related to statistics while *relevance/relevant* should be related to biology

- "EFSA Experts and Staff should be encouraged to use the interpretation of *biological relevance* and the definition of *statistical significance* specified in this document

- "Hypothesis testing should not be used as the sole tool for decision making and the level of statistical significance should not be used as the main driver to derive conclusions

- "If *statistical significance* is reported it should always be reported together with the specific statistical test used, sample sizes and the size of the effect detected and then the actual probability (*P*) values should be given

- "Results of statistical testing should not be dichotomised into significant and not significant. If, however, the results have been described as "not significant", the study design should be explored to see whether it had sufficient statistical power to detect biologically relevant effects

- "Appropriate correction methods should be considered when dealing with multiple testing. If multiple comparison methods are used in the analysis, these should be unambiguously defined. It should be clear from the text or legends to tables/figures if the *P*-values reported have been adjusted to account for multiple comparisons

- "The raw data, the programming code and all associated outputs (e.g. results and logs) from the statistical analysis should be provided to the assessor in electronic form

- "The assumptions underlying the analysis should be tested and alternative analyses should be presented/investigated to study the robustness of any results

- "Retrospective power analysis should not be conducted

- "Less emphasis should be placed on the reporting of statistical significance and more on statistical point estimation and associated confidence intervals."

Biological relevance

21. In 2017, EFSA published 'Guidance on the assessment of the biological relevance of data in scientific assessments' (EFSA 2017b)[4]. This document develops a framework and decision tree for establishing biological relevance and includes example case studies.

22. The qualitative response(s) of a biological system to an exposure (the nature of an effect) may be adverse, adaptive or beneficial, and may occur at different biological levels (e.g. molecule, cell, tissue, organ). Adverse effects may be primary (directly induced by the exposure) or secondary (e.g. related to other processes that are induced) and may be reversible or irreversible. The level of ability to absorb disturbance before a system change or loss of normal function occurs may be termed 'resilience'. This homeostatic capacity of biological systems can be variable. Adaptive effects can allow a cell or organism to survive in a changed environment without impairment of function. This may be a homeostatic response that maintains a parameter within a normal physiological range, or it may comprise a response outside of normal physiological boundaries that may eventually become adverse. Beneficial effects are alterations that lead to an improved health outcome; evaluation of such effects usually requires them to be demonstrated directly in the organism of interest rather than in a surrogate (e.g. in humans rather than an animal model).

23. The narrative notes that when an agent causes an adverse effect in an organism, the effect is often a result of a sequence of events starting with a molecular interaction between the agent and the organism. Concepts of mechanism of action, mode of action (MoA) and adverse outcome pathways (AOP) are discussed. MoA is defined by the World Health Organisation (WHO) as 'a biologically plausible sequence of key events leading to an observed effect, supported by robust experimental observations and mechanistic data'. Key cytological and biochemical events within the MoA leading to an effect are necessary and should be measurable, and magnitude of effect may be the defining factor in the determination of biological relevance. Mechanism of action is defined by WHO/ International Programme on Chemical Safety (IPCS) as the specific biochemical interaction through which a substance produces an effect on a living organism or in a biological system. MoA information can be used to establish an AOP, which indicates causal links between a molecular initiating event (MIE), intermediate key events (KE) and an adverse outcome (AO). The EFSA narrative notes that the concept of MoA could also be applied to beneficial effects of an agent; for example, the establishment of dietary reference values for food constituents.

24. A threshold (effect threshold) is defined by WHO as a dose or exposure concentration of an agent below which a stated effect is not observed or expected to occur, and a threshold dose as the dose at which an effect just begins to occur. The threshold dose can vary for a chemical depending on the effect, and also between individuals and within individuals over time. The concept of threshold doses is

[4] https://efsa.onlinelibrary.wiley.com/doi/10.2903/j.efsa.2017.4970 (accessed November 2020)

discussed further, in particular the concept that a 'true' threshold dose for a chemical or individual may not exist; as dose decreases, the dose-response curve approximates the background response and effects within the dose range become experimentally non-observable. These issues will be impacted by study design, including the power of the study to detect effects. A biological threshold does not necessarily mean that the response below this is zero, but that it may be considered to be biologically irrelevant provided the study is sufficiently powered.

25. Critical effect and critical effect size are discussed. It is re-emphasised that statistical significance does not equate to an important, meaningful, or biologically relevant outcome. Similarly, lack of statistical significance should not be taken as justification to conclude the absence of an exposure-related effect. An example is given whereby a statistically significant increase in effect is seen with 'dose 1' exposure to an agent in comparison with control (no exposure). However, the effect size is within the known background variability and thus the 'dose 1' effect would not be considered meaningful, while a greater effect size at (higher) 'dose 2', above known background variability, would be taken as the lowest observed effect level (LOEL) for the study. Furthermore, an effect of magnitude outside control variability may still not be biologically relevant: in order to determine biological relevance, it should be considered whether the effect could actually lead to functional deficit later in the study[5].

26. The EFSA Scientific Committee developed a framework for biological relevance, comprising three main stages: development of an assessment strategy (specification of agents, effects, subjects and conditions); collection and extraction of relevant data (identification of potentially biologically relevant evidence/data as specified in the assessment strategy); appraisal and integration of the relevance of the agents, subjects, effects, and conditions.

Assessment strategy

27. The aim of developing the assessment strategy is to define the protocol for data collection, to ensure that the assessment will answer the question(s) posed. Aspects to be considered include: specification of agents of interest; subjects or populations to be covered by the assessment; effects of exposure that would be considered relevant to the assessment question; and conditions of exposure that are relevant to the assessment question, such as route, timing and duration. A main objective is to identify and specify biologically relevant data before data collection is initiated. Standardised procedures (e.g. guidance documents) may already cover such aspects.

Collection and selection of data

28. All data should be collected and considered, and criteria for subsequent inclusion/exclusion should be described. Information should be evaluated for

[5] This discussion/example relates to reproductive toxicity studies.

relevance to the question posed; this aspect is considered in more detail in the EFSA Scientific Opinion 'Guidance on the use of the weight of evidence approach in scientific assessments' (EFSA 2017a).

Appraisal and integration of the data

29. Data should be considered for relevance to the assessment questions, as defined in the assessment strategy. Causal relationship of exposure and effect can be assessed by referral to Bradford-Hill considerations: is the effect dose-related; is there confounding; does exposure precede effect on a plausible timescale; is the effect biologically plausible; is there information on MoA? Subsequently, it is necessary to determine whether the nature and size of the effect is relevant to the assessment question. A number of questions are proposed to aid in this evaluation. These address aspects such as whether the effect itself is an adverse or beneficial effect or is linked to an adverse/beneficial outcome, and for any of these situations, is the effect size of sufficient magnitude to be considered relevant? Equivalence testing may be helpful to identify values that fall outside of normal, natural variation, and furthermore may help in concluding on the relevance of the effects in terms of safety. A critical effect size can be determined by expert judgement. In cases where a consensus cannot be reached, the EFSA Scientific Committee recommend that default values should be used – a critical effect size or benchmark response (BMR) of 10% (extra risk) for quantal data and 5% (change in mean response) for continuous data from animal studies. The rationale for deviating from or using default values should be documented.

30. A decision tree is presented to aid decision in whether a biological effect is relevant or not, reproduced in Figure 1 below.

31. Relevance of the test subject should be taken into account. For animal studies, this can include judgement based on the MoA of generation of an effect, if this is known. Qualitative and quantitative interspecies differences should be taken into account, including toxicokinetic (TK) and toxicodynamic (TD) processes.

32. The inclusion of evidence with less biological relevance increases the overall uncertainty in an assessment. Uncertainties in biological relevance should be addressed and described at all stages of an evaluation, along with other uncertainties and data gaps. Methods for assessing uncertainty have been addressed in the EFSA Guidance Document 'Guidance on Uncertainty Analysis in Scientific Assessments' (EFSA 2018).

33. In its conclusions, the EFSA Scientific Committee notes that, in the broad sense, the concept of biological relevance in risk assessment encompasses aspects relevant to problem formulation as well as relating to the narrower interpretation of biological relevance of an effect. Relevance is a fundamental concept in dealing with evidence and has different implications at different stages of an analysis. These can only be determined when the question is well defined.

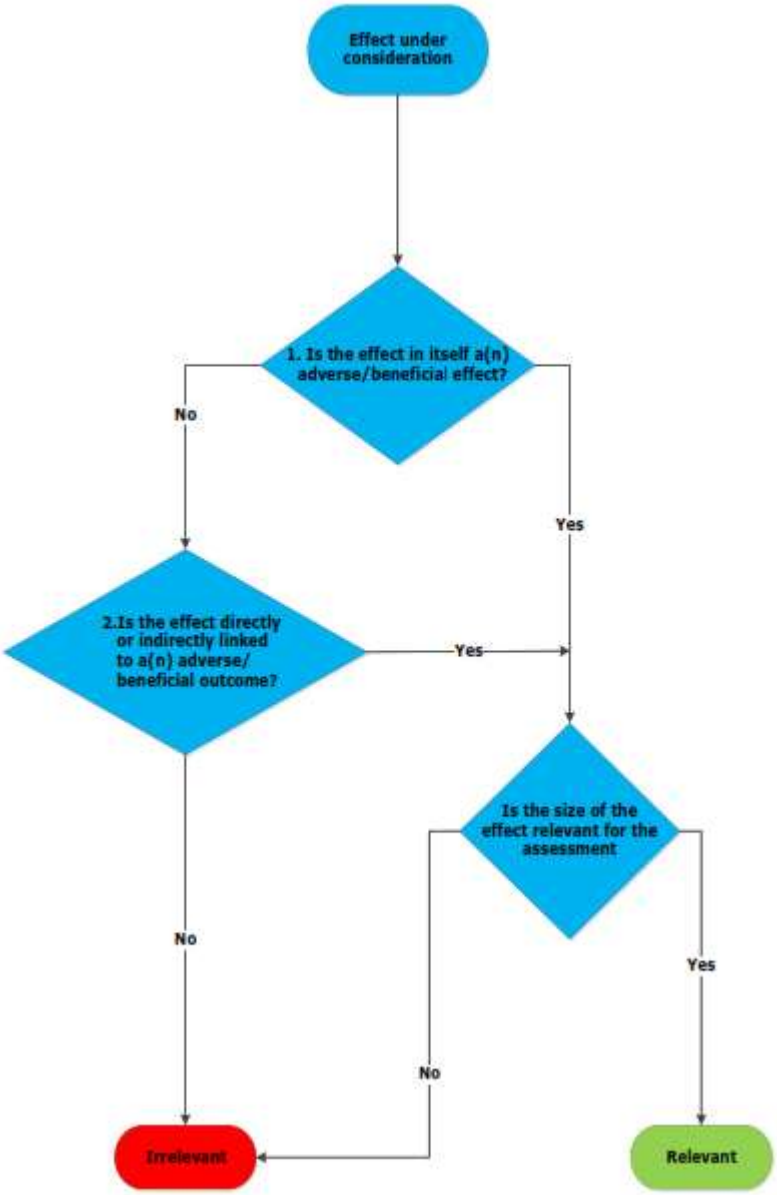

Figure 1. General decision tree to decide whether a biological effect is relevant or not. *Reproduced from Figure 5 of* EFSA 2017b.

Statistical approaches and their limitations

34. Lovell (2013) published a commentary entitled 'Biological importance and statistical significance' which explored statistical ideas behind the analysis of experiments related to crop composition and the genetic factors underlying composition, drawing on work carried out by the EFSA Statistical Working Group that led to the EFSA (2011) opinion discussed in paragraphs 9-20, above. In this publication, the author emphasises the particular importance of good experimental design for subsequent adequate statistical analysis of the data set. Although null hypothesis statistical testing (NHST), with the identification of *P*-values and statistical significance, appear to be the primary objective of the majority of analyses, it would be better to place emphasis on the identification of the size of effects that are

biologically important. To achieve this adequately, involvement of scientists with in-depth knowledge in the domain of interest is necessary at the study planning stage.

35. The narrative covers aspects of experimental design, hypothesis testing and statistical significance, criticism of the use of significance, biological importance over statistical significance, alternative use of confidence intervals and estimation, equivalence testing, multiple comparisons, modelling, multivariate and graphical methods, and Bayesian approaches.

36. In terms of experimental design, the author notes that studies in agricultural sciences often work via a factorial 'design of experiment' (DOE) approach, rather than 'one factor at a time' (OFAT). However, DOE and its advantages have generally not been extended to other domains of study.

37. The distinction is highlighted between the Fisherian approach of statistical significance testing and the Neyman-Pearson approach of hypothesis testing. Neyman-Pearson is a binary decision between two hypotheses, while Fisherian evaluation establishes a *P*-value that is used to decide whether or not the null hypothesis is rejected. NHST is a hybrid of these two methodologies. In NHST, the test statistic and *P*-value are affected by factors including sample size, statistical test used, and amount of variability. The size of difference that is just significant (reaches the critical value of the test statistic) will vary from study to study. Each experiment is one of a range of possible experiments and thus gives an estimate and distribution of the true difference. Thus, it is possible that an individual experiment may produce an estimate within a tail and the study would be reported as not significant even if there was a real difference (Type II error).

38. The narrative criticises over-reliance on the concept that 'statistical significance is synonymous with $P<0.05$', noting that the *P*-value represents the probability of obtaining the data set by chance alone, but *not* the probability that the null hypothesis is true. The commentary also criticises the common confusion between 'significance' and 'importance'. In this respect, $P<0.05$ is seen as a definitive requirement for acceptance for journal publication of findings, which can further entrench the over-reliance on the use of *P*-values. Additionally, as also emphasised by EFSA, the *P*-value does not give an indication of effect size; as such many statisticians propose that reporting of *P*-values should be replaced by other methods (e.g. estimates and CIs; use of Bayesian statistical methods).

39. Lovell (2013) reaffirms the opinion of EFSA that biological importance should take precedence over statistical significance. Defining what is biologically relevant is not a statistical decision but has important implications for study design and for subsequent statistical analysis and interpretation of findings. One concept in study design is the 'minimal difference that you can afford to miss'. This can be equated to the concept of 'clinically relevant difference' (CRD) in clinical studies, where findings are sometimes categorised into standardised effect sizes (small, medium, large). In toxicological studies, the choice of effect size would be a decision for the expert

scientist in the relevant study domain. Experiments designed with power for a primary endpoint may have higher or lower power for secondary endpoints; in considering this, the existence of historical information in reference databases such as the US Environmental Protection Agency (EPA) ToxRefDB can help by providing information on estimates of variability and size of effects that would be expected under specific experimental designs.

40. The concept of equivalence may refer to either 'substantial equivalence' or 'bioequivalence'. Substantial equivalence (for example, equivalence of novel foods such as genetically modified foods) is a concept developed by OECD and is important in a regulatory perspective. Bioequivalence is a pharmaceutical concept and aims to ensure that products are not declared to be equivalent simply through lack of adequate capability of a clinical study to detect a difference. Bioequivalence testing was developed to overcome problems associated with NHST and has been extended to concepts of non-inferiority and superiority tests[6]. Acceptable intervals for sample size and power calculations (Δ) are pre-defined in the study protocol.

41. When multiple comparisons are being made (in toxicological studies, for example body weight, organ weight, clinical chemistry, urine analysis and haematology), using an NHST approach there is a high likelihood of 'statistically significant' findings being observed by chance. Methods for multiple comparison can be used to avoid this (e.g. Bonferroni correction, Dunnett's test). Many different tests exist, with a wide variation in the degree to which they impact study outcomes. There is some concern among statisticians that journals may be over-prescriptive in the use of specific tests, which may not always be appropriate to the study conducted.

42. Finally, the paper discusses modelling, multivariate and graphical methods, and Bayesian methods. These detailed aspects are outside the scope of the present document.

43. More recently, Lovell (2020) published a paper entitled 'Null hypothesis significance testing and effect sizes: can we 'effect' everything… or …anything?'. This publication addresses, develops, and updates some of the issues in statistical design and analysis raised by Lovell (2013), with a perspective towards studies in pharmacology, psychology, and epidemiology.

44. Lovell (2020) notes the recent publication of an article in the journal, *Nature*, entitled 'Retire statistical significance', which was signed by more than 800 statisticians. This publication advocated '…the entire concept of statistical significance to be abandoned', including '…a stop to the use of *P*-values in the conventional, dichotomous way – to decide whether a result refutes or supports a scientific hypothesis' (Amrhein et al 2019, *cited in* Lovell (2020)). Concurrently, the American Statistical Association (ASA) has addressed this issue with an extensive

[6] Tests to evaluate whether a new intervention is as good as or better than the standard intervention. For non-inferiority, new ≥ standard; for superiority, new > standard; for equivalence, new/standard = 1 ± α.

series of papers in a 2019 issue of the journal 'American Statistician' that addresses issues including:

- The appropriateness of the traditional NHST paradigm
- Whether identifying the size and biological importance of an effect is more important than whether the difference attains a threshold such as $P<0.05$
- The use of effect sizes to help interpret and design studies
- Whether the development of Bayesian statistical methods represents a realistic challenge to the more traditional frequentist approaches.

45. Many statisticians have problems with the NHST approach, and some have advocated banning *P*-values. However, others argue that there are areas of research where a binary approach is useful (e.g. genome-wide association studies, quality control). In addition, there is a concern that abandonment of NHST may lead to a less acceptable situation in which statistical analysis is replaced by subjective assessment; abandonment of $P=0.05$ as a 'gatekeeper' may allow researchers to fit findings to a pre-existing narrative. Others have expressed a preference to focus on effect sizes and/or alternative approaches such as Bayesian methods, or perhaps to lower the *P*-value cut-off to 0.005. It is noted that, in fact, the use of estimates and CIs has been accepted since 1988, but there is a problem in the definition of a CI, which may still be underpinned by the concept of NHST. A 'credibility interval' has been proposed as an alternative, whereby the estimated parameter is treated as a random variable with fixed boundaries (the converse of a confidence interval). Overall however, a general theme has emerged that estimates and CIs are better than NHST.

46. The concept of effect sizes may be applied in two ways: observed effect sizes (generally broad) or planned effect sizes (generally narrow). In replacing NHST with effect sizes, an 'effect size movement' has developed in the field of educational and psychology research, with some journals mandating the reporting of effect sizes. There is criticism of this approach, and it is argued strongly that interval estimates should accompany effect sizes. The effect size in consideration should be relevant to the particular research question being addressed.

47. The use of standardised effect sizes has the advantage that they can be compared across studies. Effect statistics and CIs are an absolute requirement for meta-analysis. In addition, unstandardised data should also be presented to allow for calculation of standardised results when a meta-analysis is carried out.

48. There is an increasing practical capability to use, and an acceptance of, Bayesian approaches, and this has led to alternative approaches to statistical testing. However, these approaches also have some critics.

49. Lovell (2020) concludes that NHST is accepted to have many limitations. It is so widely used that there is fear that attempts to replace it will result in a less acceptable situation – where decisions are made simply using subjective judgement. Estimation approaches such as point estimates and CIs provide an alternative approach giving information on effect size and uncertainty, but still have some limitations. Approaches using effect sizes are useful in study design and in meta-analysis. However, the use of effect size to assess results in the absence of limits such as CIs is not good practice. Alternative methods are now being suggested, including Bayesian methods, and these issues were discussed in an ASA special issue, published in 2019[7]. Lovell (2020) comments that no single method is likely to provide an alternative to NHST, but statisticians should continue to educate researchers, authors, reviewers and editors on inappropriate use of NHST. It should be appreciated that objective use of statistical analysis is not to provide certainty to a decision, but rather bounds on the degree of uncertainty. The role of statistics is to provide estimates of effect sizes and a degree of measure of uncertainty, not a binary significant/nonsignificant, positive/negative conclusion: scientists should learn more about the subtleties of experimental design, statistical methods, and interpretation of results in addition to the core skills of using statistical packages.

Summary

50. The problem of determining how experimental results can best be judged to establish their importance has led to a debate relating to the relationship between biological relevance and statistical significance. The judgment of statistical significance is typically based on an assessment of the likelihood of the observed effect occurring by chance alone; when a result has a less than 5% chance of occurring by chance ($P<0.05$) it is usually judged as statistically significant. Such categorisation of results as either ‘significant’ or ‘not significant’ is often mistakenly taken as an indication of mathematical proof of a biologically relevant effect (or lack of), although in reality the *P*-value cut-off gives no more information than how likely the data are to have occurred by chance.

51. Although the use of NHST is criticised by many statisticians, there is also a concern that attempts to replace this methodology could result in a less-acceptable situation, where the importance of study findings is judged subjectively. The calculation of a biological effect using a statistical point estimate and its uncertainty (interval estimate; confidence intervals (CIs)) is an alternative approach that provides more information than the simple result of a significance test. Alternatives such as Bayesian methods are also proposed.

52. Identifying statistical significance should not be the main objective of a statistical analysis of study data. The focus should be on identifying sizes of effects that are biologically important. The involvement of expert scientists within the domain of interest is critical from the planning stage, with an aim to design studies with

[7] Am Statistician 2019, 73.

sufficient statistical power. Ultimately, biological relevance should be the primary factor of importance.

Initial comments on aspects of this paper

53. A number of points are raised about study power and study design. One difficulty is that scientific advisory committees almost always have to assess data that has already been generated, so any guidance is needed on both study design for those who are generating data and study interpretation, for those, such as COM, COC and COT, who are assessing the data. In addition, Committees examining study data may seek to undertake retrospective power analyses to determine if a study was designed to detect a biologically relevant effect (see paragraph 15 and bullet points in paragraph 20).

54. With respect to the relative importance of biological relevance and statistical significance (see paragraph 16), there is also the question of whether an intervention did produce an effect, if there is not statistical significance. There may be a danger of conflating scientific assessment of a study with the precautionary principle. The limits of a study and its uncertainty need to be stated; it is then a policy decision as to how to address this, with responsibility varying by jurisdiction.

55. In contrast, where statistical significance is identified for an effect that is of a size that is not biologically relevant, this could be taken into account in designing the test strategy, for example by testing for a minimum change in effect that is considered biologically relevant (i.e. an adverse effect) rather than for no effect (see paragraph 18, 29 and 52). It should also be noted that there is ‘noise’ (uncertainty) in all measurements. Even in the absence of any true effect, a confidence interval with a positive upper bound (i.e. indicative of a possible effect of a given magnitude) is possible. As such there is a need for mechanistic information and weight-of-evidence integration when considering the effects of exposure to a substance.

56. A challenge in taking forward use of confidence intervals (see paragraph 19), is when looking at data where a NOAEL approach has been used, as this does not provide CIs, though these can be calculated but with difficulty, and only when all of the data are available.

57. Finally, some of the papers outlined, have also flagged aspects relating to beneficial effects as well as risks of exposures. This is outside the remit of the COC, COM and COT, however it is noted that risk-benefit is complex, and considerations of whether benefit needs to be demonstrated in the target species whereas hazard can be determined in surrogate species, is a policy decision based on weighting of uncertainties, but scientifically it should be judged on the biology (see paragraph 22).

Questions for the Committees

58. Members are asked to consider this scoping paper and comment on the aspects covered, and how they wish this topic to be taken forward across the three Committees.

59. A list of further papers is provided after the reference list that can be considered as the topic is taken further, and Members are invited to provide any additional references that would be relevant.

NCET at WRc/IEH-C under contract supporting the PHE COT Secretariat
November 2020

Abbreviations

AOP	Adverse Outcome Pathway
ASA	American Statistical Association
BMR	Benchmark Response
CI	Confidence Interval
CRD	Clinically Relevant Difference
DOE	Design of Experiment
EFSA	European Food Safety Authority
EPA	US Environmental Protection Agency
IPCS	International Programme on Chemical Safety
KE	Key Event
LOEL	Lowest Observed Effect Level
MIE	Molecular Initiating Event
MoA	Mode of Action
NHST	Null Hypothesis Statistical Testing
OECD	Organisation for Economic Co-operation and Development
OFAT	One Factor at a Time
SD	Standard Deviation
TD	Toxicodynamic
TK	Toxicokinetic
WHO	World Health Organisation

References

EFSA. 2011. EFSA Scientific Committee; Statistical Significance and Biological Relevance. EFSA Journal 2011;9(9):2372 [17pp.] https://www.efsa.europa.eu/en/efsajournal/pub/2372.

EFSA 2017a. EFSA Scientific Committee, Hardy A, Benford D, Halldorsson T, Jeger MJ, Knutsen HK, More S, Naegeli H, Noteborn H, Ockleford C, Ricci A, Rychen G, Schlatter JR, Silano V, Solecki R, Turck D, Benfenati E, Chaudhry QM, Craig P, Frampton G, Greiner M, Hart A, Hogstrand C, Lambre C, Luttik R, Makowski D, Siani A, Wahlstroem H, Aguilera J, Dorne J-L, Fernandez Dumont A, Hempen M, Valtue~na Martınez S, Martino L, Smeraldi C, Terron A, Georgiadis N and Younes M, 2017. Scientific Opinion on the guidance on the use of the weight of evidence approach in scientific assessments. EFSA Journal 2017;15(8):4971, 69 pp. https://doi.org/10.2903/j.efsa.2017.4971 https://www.efsa.europa.eu/en/efsajournal/pub/4971.

EFSA. 2017b. EFSA Scientific Committee, Hardy A, Benford D, Halldorsson T, Jeger MJ, Knutsen HK, More S, Naegeli H, Noteborn H, Ockleford C, Ricci A, Rychen G, Schlatter JR, Silano V, Solecki R, Turck D, Younes M, Bresson J-L, Griffin J, Hougaard Benekou S, van Loveren H, Luttik R, Messean A, Penninks A, Ru G, Stegeman JA, van der Werf W, Westendorf J, Woutersen RA, Barizzone F, Bottex B, Lanzoni A, Georgiadis N and Alexander J, 2017. Guidance on the assessment of the biological relevance of data in scientific assessments. EFSA Journal 2017;15(8):4970, 73 pp. https://doi.org/10. 2903/j.efsa.2017.4970 https://www.efsa.europa.eu/en/efsajournal/pub/4970.

EFSA. 2018. EFSA (European Food Safety Authority) Scientific Committee, Benford D, Halldorsson T, Jeger MJ, Knutsen HK, More S, Naegeli H, Noteborn H, Ockleford C, Ricci A, Rychen G, Schlatter JR, Silano V, Solecki R, Turck D, Younes M, Craig P, Hart A, Von Goetz N, Koutsoumanis K, Mortensen A, Ossendorp B, Martino L, Merten C, Mosbach-Schulz O and Hardy A, 2018. Guidance on Uncertainty Analysis in Scientific Assessments. EFSA Journal 2018;16(1):5123, 39 pp. https://doi.org/10.2903/j.efsa.2018.5123 https://www.efsa.europa.eu/en/efsajournal/pub/5123.

Lovell, D. P. (2013) Biological importance and statistical significance. *J Agric Food Chem,* 61**,** 8340-8.

Lovell, D. P. (2020) Null hypothesis significance testing and effect sizes: can we 'effect' everything … or … anything? *Curr Opin Pharmacol*.

Further Reading

ECETOC (2002). Recognition of, and Differentiation between, Adverse and Non-adverse Effects in Toxicology Studies. http://www.ecetoc.org/wp-content/uploads/2014/08/ECETOC-TR-085.pdf.

Felter et al (2020). Hazard identification, classification, and risk assessment of carcinogens: too much or too little?–Report of an ECETOC workshop. Critical Reviews in Toxicology, 50(1), 72-95. https://www.tandfonline.com/doi/full/10.1080/10408444.2020.1727843

Lee (2016). 'Alternatives to P value: confidence interval and effect size.' Korean J Anesthesiol, 69(6), 555-562. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5133225/

Mellis et al (2018). 'Lies, damned lies and statistics: Clinical importance versus statistical significance in research. Paediatric Respiratory Reviews, 25, 88-93. https://www.sciencedirect.com/science/article/abs/pii/S1526054217300088?via%3Dihub (abstract).

Pandiri et al (2017) 'Is it adverse, non-adverse, adaptive or artifact?'. Proceedings of an SOT continuing education course on differentiating adverse effects from non-adverse of adaptive effects. Toxicol Pathol, 45(1), 238-247. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5225139/

Schmidt et al (2016). 'Enhancing the interpretation of statistical P values in toxicology studies: implementation of linear mixed models (LMMs) and standardized effect sizes (SESs)'. Arch Toxicol, 90, 731-751. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4754325/

Schober et al (2018). Statistical significance versus clinical importance of observed effect sizes: what do *P* values and confidence intervals really represent? Anesth Analg, 126(3), 1068-1072. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5811238/pdf/ane-126-1068.pdf

Szucs and Ioannidis (2017). 'When null hypothesis significance testing is unsuitable for research: a reassessment.' Front Hum Neurosci, 11, 390. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5540883/