Organisation de Coopération et de Développement Économiques
Organisation for Economic Co-operation and Development 20-Jul-2017

English - Or. English

ENVIRONMENT DIRECTORATE
JOINT MEETING OF THE CHEMICALS COMMITTEE AND THE WORKING PARTY ON CHEMICALS, PESTICIDES AND BIOTECHNOLOGY

Background Review Document Supporting the Development of the Test Guideline 433 on Acute Inhalation Toxicity – Fixed Concentration Procedure Harmonised Submission

Series on Testing & Assessment
No. 265

JT03417499

ENV/JM/MONO(2017)17
Unclassified
English - Or. English

OECD Environment, Health and Safety Publications

Series on Testing and Assessment

No. 265

BACKGROUND REVIEW DOCUMENT SUPPORTING THE DEVELOPMENT OF THE TEST GUIDELINE 433 ON ACUTE INHALATION TOXICITY - FIXED CONCENTRATION PROCEDURE HARMONISED SUBMISSION

IOMC INTER-ORGANIZATION PROGRAMME FOR THE SOUND MANAGEMENT OF CHEMICALS

A cooperative agreement among FAO, ILO, UNDP, UNEP, UNIDO, UNITAR, WHO, World Bank and OECD

Environment Directorate
ORGANISATION FOR ECONOMIC CO-OPERATION AND DEVELOPMENT
Paris 2017

About the OECD

The Organisation for Economic Co-operation and Development (OECD) is an intergovernmental organisation in which representatives of 35 industrialised countries in North and South America, Europe and the Asia and Pacific region, as well as the European Commission, meet to co-ordinate and harmonise policies, discuss issues of mutual concern, and work together to respond to international problems. Most of the OECD's work is carried out by more than 200 specialised committees and working groups composed of member country delegates. Observers from several countries with special status at the OECD, and from interested international organisations, attend many of the OECD's workshops and other meetings. Committees and working groups are served by the OECD Secretariat, located in Paris, France, which is organised into directorates and divisions.

The Environment, Health and Safety Division publishes free-of-charge documents in twelve different series: **Testing and Assessment**; **Good Laboratory Practice and Compliance Monitoring**; **Pesticides**; **Biocides**; **Risk Management**; **Harmonisation of Regulatory Oversight in Biotechnology**; **Safety of Novel Foods and Feeds; Chemical Accidents; Pollutant Release and Transfer Registers; Emission Scenario Documents; Safety of Manufactured Nanomaterials;** and **Adverse Outcome Pathways.** More information about the Environment, Health and Safety Programme and EHS publications is available on the OECD's World Wide Web site (www.oecd.org/chemicalsafety/).

This publication was developed in the IOMC context. The contents do not necessarily reflect the views or stated policies of individual IOMC Participating Organizations.

The Inter-Organisation Programme for the Sound Management of Chemicals (IOMC) was established in 1995 following recommendations made by the 1992 UN Conference on Environment and Development to strengthen co-operation and increase international co-ordination in the field of chemical safety. The Participating Organisations are FAO, ILO, UNDP, UNEP, UNIDO, UNITAR, WHO, World Bank and OECD. The purpose of the IOMC is to promote co-ordination of the policies and activities pursued by the Participating Organisations, jointly or separately, to achieve the sound management of chemicals in relation to human health and the environment.

This publication is available electronically, at no charge.

For this and many other Environment, Health and Safety publications, consult the OECD's World Wide Web site (www.oecd.org/ehs)

or contact:

OECD Environment Directorate,
Environment, Health and Safety Division
2, rue André-Pascal
75775 Paris cedex 16
France

Fax : (33-1) 44 30 61 80

E-mail : ehscont@oecd.org

FOREWORD

This document contains background information and relevant references supporting issues raised with the Fixed Concentration Procedure test method for acute inhalation toxicity in OECD Test Guideline 433 (TG 433). It was prepared by the United Kingdom who has led the development of TG 433.

Part 1 of the Background Document presents information available in April 2016 for the 28th Meeting of the Working Group of the National Coordinators of the Test Guidelines Programme (WNT-28) where the draft TG 433 was discussed for approval.

Part 2 of the Background Document presents the outcome of the analysis and complementary work performed after WNT-28 in 2016 to address issues identified.

The Background Document was approved by the Working Group of the National Co-ordinators of the Test Guidelines Programme (WNT) at its 29th meeting in April 2017, together with the TG 433. The Joint Meeting of the Chemicals Committee and the Working Party on Chemicals, Pesticides and Biotechnology agreed to the declassification of he Background Document on 10th July, 2017. This document is published under the responsibility of the Joint Meeting of the Chemicals Committee and the Working Party on Chemicals, Pesticides and Biotechnology.

BACKGROUND INFORMATION:
PART 1 (APRIL 2016)

REVISED OECD TG 433 (FIXED CONCENTRATION PROCEDURE)

Background

1. The Fixed Concentration Procedure (FCP) in acute inhalation studies of chemicals (OECD TG433) uses fewer animals than current test guidelines and does not require lethality as an endpoint. However, this was dropped from the OECD work plan in 2007 due to three main areas of concern:

1. A lack of evidence for comparable performance with TG403 and TG436;

2. Suspected sex differences in the level of toxic effects (since the FCP was originally proposed to use females as the default sex);

3. The ill-defined and subjective nature of evident toxicity.

2. In 2008 the NC3Rs began a three-stage strategy of work in collaboration with inhalation toxicity experts to address these issues. These issues have now been resolved (as outlined below) and have been used to revise the draft Test Guideline 433. These have been submitted for consideration at the April 2016 WNT meeting.

Comparison to existing methods and potential for sex differences

3. The original protocol: The original test method tested proposed using females only, for both the sighting and main study, unless males were indicated to be more sensitive. Previous work of the group used a statistical simulation approach to show that the original TG433 procedure classified substances either in the same or occasionally a more stringent GHS class than that assigned on the basis of the LC_{50} value (Stallard *et al*., 2003). The group then commissioned a statistical evaluation to compare the three test methods, which showed that all three methods perform well in the absence of sex differences. However, performance is affected in ***all*** cases in the presence of unanticipated sex differences (Price *et al*., 2011). With TG403, a sex difference leads to a slightly greater chance of under-classification. This is also the case for TG436, but more pronounced than for TG403. For the original TG433 draft protocol (which proposed the use of females only as the default sex), the classification is unchanged if females are more sensitive. However, if males are more sensitive, the procedure may lead to under-classification. This issue was addressed by a modification of the sighting study in the revised TG433 to first test both a male and female animal to identify potential differences in sensitivity.

4. The revised protocol with modified sighting study: Statistical evaluation of the performance of the revised protocol showed that sex differences in sensitivity do not significantly impact on the performance of the FCP; testing of one male and one female is statistically viable and the addition of this sighting study makes TG433 more robust (Stallard *et al*., 2011). This work has been published in *Human and Experimental Toxicology* (Price *et al*., 2011; Stallard *et al*., 2011).

5. Following recent comments questioning the validity of testing only one male and one female, Stallard conducted further statistical simulations to show that where there is more than 10-fold difference in sensitivity between sexes, the probability of choosing the most sensitive sex in the modified sighting study is almost 100% (Appendix I). Sex differences smaller than this are unlikely to impact classification in the main study. Furthermore, sex differences are uncommon in inhalation studies. The publication by Price *et al.* (2010) showed that a statistically significant sex difference in 16 of 56 studies (29%), where females were the more sensitive sex in 11/16 (69%) studies. No statistically significant sex differences were identified in the dataset by Sewell *et al*. (2015).

6. A sighting study with single animals has been found appropriate for the acute oral equivalent of TG433, the fixed dose procedure (TG420). This test guideline uses females as default for both the sighting and the main studies, unless prior information has indicated that males are more sensitive. Though prior information can also be used in TG433 to aid the choice of the most appropriate sex, the inclusion of both males and females in the sighting study gives an additional opportunity to identify any large sex differences.

7. However, the main purpose of the sighting study is to identify an appropriate starting dose for the main study, as testing at an inappropriate starting dose can result in unnecessary studies in animals. The use of a sighting study such as this offers an advantage over the accepted TG436 which does not include a sighting study, and therefore has a higher potential for an inappropriate starting dose to be used resulting in unnecessary suffering/testing in animals.

Evident toxicity

8. Using the FCP, a decision on whether further testing is required at higher and potentially more toxic concentrations is made by assessing the animals for signs of toxicity, termed ‘evident toxicity’, rather than relying on severe toxicity or deaths. The concept of ‘evident toxicity’ has been successfully established over many years by the conduct of the Acute Oral FDP (TG420) for regulatory submissions. The original version of TG420 was adopted in 1992 after the reproducibility of the procedure was established in large scale *in vivo* validation trials (van den Heuvel *et al*., 1990). Nevertheless, some members of the OECD inhalation expert working group have expressed concerns that the definition of ‘evident toxicity’ is insufficiently clear. We have collected data to build an evidence-base to address this concern, in order to provide objective guidance on the recognition of evident toxicity and make the decision less subjective. Data on the clinical signs observed in individual animals during acute inhalation studies has been collected for 188 substances from 511 acute inhalation studies provided by seven laboratories worldwide (US, EU, Korea and Japan). These have been analysed to determine if there are any signs observed at the lower dose that could have predicted severe toxicity or death at the higher concentration. Signs such as body weight loss (>10% pre-dosing weight), irregular respiration, tremors and hypoactivity, seen at least once in at least one animal after the day of dosing are highly predictive (positive predictive value >90%) of severe toxicity or death at the next highest concentration. This has been published in *Regulatory Toxicology and Pharmacology* (Sewell *et al*., 2015). However, confidence that evident toxicity has been reached will of course grow with increasing number of animals displaying the sign, and if more than one of the four signs is displayed.

9. One of the original concerns raised was that it would be difficult to observe for signs of toxicity during the exposure period itself, particularly if the nose-only chamber was used (as preferred for TG433 and TG436). However, our guidance for recognition of evident toxicity is based on the presence of the clinical signs that occur after the day of exposure. This was not only to avoid inclusion of signs associated with the dosing or restraint procedure, but also to take into account and select for signs severe enough to persist to the next day.

10. The analysis conducted by Sewell *et al*., (2015) was based on a minimum two-fold concentration change, to be representative of the fold changes that might reasonably be considered between concentrations in the existing OECD TGs. If it is predicted that animals would die or suffer severe toxicity at a two-fold higher concentration, then it would also follow that the same would apply to even larger increases in concentration. Therefore by using a fold change at the lower end of what would be used in practice, the predictive ability of the signs would be expected to be *greater* when fold changes are greater. Similarly, wrongly predicting toxicity at the next does is *less* likely when fold change is greater.

11. It is important to note that 'evident toxicity' is already an accepted endpoint used in TG420, which provides no additional guidance on its recognition. TG433 uses this same endpoint, but provides additional guidance to aid the decision on whether evident toxicity has been reached. Sewell *et al*. (2015) showed that the four signs identified, as well as being highly predictive, were also highly sensitive and specific. They successfully predicted a large proportion of the total toxicities in the dataset and false positive/negatives were low. The publication lists other rarer (therefore less sensitive due to wider confidence intervals) signs that were also highly predictive and may indicate that evident toxicity has been reached. This information is intended to guide the decision on whether evident toxicity has been reached, complementing study director judgment and experience. However, as a last resort (e.g. in situations where the signs displayed make the decision of whether evident toxicity has been reached unclear), TG433 also allows for classification to be determined on the same basis as TG436, i.e. using death or euthanasia alone (outcome A).

12. The use of evident toxicity rather than death as an endpoint provides a significant refinement over existing methods which use death as an endpoint. It will improve animal welfare through reduced suffering, but will also reduce the number of studies required, as classification can often be made after testing at one concentration only.

Other concerns

13. Other concerns that were raised, such as the spacing of the fixed concentrations, the use of targeted rather than actual concentrations and the requirement for pre-tests (without animals) to achieve specific concentrations and particle size, also apply to TG436 or relate to the conduct of acute inhalation studies in general.

The revised test guideline

14. The revised TG433 incorporates a modification of the sighting study to test both a male and female animal to identify and overcome the potential influence of sex differences and provides a more objective definition of evident toxicity. TG433 offers a significant refinement to existing methods by using evident toxicity rather than death as an endpoint. It also uses fewer animals. Statistical evaluation supports the adoption of this method for classification and labeling purposes.

References

Price, C., N. Stallard, S. Creton, I. Indans, R. Guest, D. Griffiths, and P. Edwards. 2011. A statistical evaluation of the effects of gender differences in assessment of acute inhalation toxicity. *Human & experimental toxicology*. 30:217-238.

Sewell, F., I. Ragan, T. Marczylo, B. Anderson, A. Braun, W. Casey, N. Dennison, D. Griffiths, R. Guest, T. Holmes, T. van Huygevoort, I. Indans, T. Kenny, H. Kojima, K. Lee, P. Prieto, P. Smith, J. Smedley, W.S. Stokes, G. Wnorowski, and G. Horgan. 2015. A global initiative to refine acute inhalation studies through the use of 'evident toxicity' as an endpoint: towards adoption of the Fixed Concentration Procedure. *Regulatory toxicology and pharmacology*. 73:770-779.

Stallard, N., C. Price, S. Creton, I. Indans, R. Guest, D. Griffiths, and P. Edwards. 2011. A new sighting study for the fixed concentration procedure to allow for gender differences. *Human & experimental toxicology*. 30:239-249.

Stallard, N., A. Whitehead, and I. Indans. 2003. Statistical evaluation of the fixed concentration procedure for acute inhalation toxicity assessment. *Human & experimental toxicology*. 22:575-585.

van den Heuvel, M.J., D.G. Clark, R.J. Fielder, P.P. Koundakjian, G.J. Oliver, D. Pelling, N.J. Tomlinson, and A.P. Walker. 1990. The international validation of a fixed-dose procedure as an alternative to the classical LD50 test. *Food and chemical toxicology : an international journal published for the British Industrial Biological Research Association*. 28:469-482.

BACKGROUND INFORMATION
PART II (MARCH 2017)

REVISED OECD TG 433 (FIXED CONCENTRATION PROCEDURE)

Background

15. Following consideration at the April 2016 WNT meeting there were still some outstanding concerns. Specifically around the:

1. Potential for TG433 to result in over classifications compared to other acute inhalation test guidelines.

2. The definition of evident toxicity – the data supporting the recommendations, the incorporation of severity and the subjective nature of evident toxicity.

3. The choice of default sex and the requirement for a sighting study.

16. Further analyses have been carried out to support this test guideline and to address the remaining concerns. These include 'validation' of comparable performance by retrospective classifications made via TG433 using the new definition of evident toxicity in the Sewell *et al*., (2015) dataset, and further analysis to support the simple definition of evident toxicity (i.e. effect of concentration ratios, number of animals displaying a sign, and signs seen in isolation versus combinations of signs).

17. The results of these further analyses are outlined below, and have been presented to and discussed with the expert inhalation toxicologists that expressed the original concerns. The group agreed that these concerns have now been addressed and final incorporations in to the test guideline have been agreed for consideration at the next WNT meeting (April 2017).

Potential for TG433 to result in over classifications – retrospective classifications

18. Previous publications have addressed comparability of the three acute inhalation methods (TG403, TG436 and TG433) using statistical simulations to compare the classifications made by each of the three methods and the likelihood of misclassification (under or over) (Price *et al*., 2011; Stallard *et al*., 2011; Stallard *et al*., 2003). The statistical simulations showed that the three methods were comparable, but that all three methods had the potential to misclassify. For the TG433 method, the statistical simulations assumed correct identification of evident toxicity to make the GHS classification predictions. We applied the recommendations for recognition of evident toxicity detailed in Sewell *et al*. (2015) (any of the following signs observed at least once in at least one animal from the day after exposure: tremors, hypoactivity, >10% bodyweight loss or irregular respiration) to the Sewell *et al*. (2015) dataset to make retrospective classifications, demonstrating how it can be used successfully in practice. There is very strong agreement (>90%) between the classifications made by each of the three methods, irrespective of the sex of animals used for TG433. For further information on the retrospective analyses see Appendix II.

The simple definition of evident toxicity

19. Outcome B in the TG433 protocol (Figure 1) is defined as 1 death and/or evident toxicity in more animal. Analysis by Sewell *et al.*, 2015 showed that the death of only 1 animal at the lower concentration is 93% predictive of toxicity (death of 2 or more animals) at the next higher concentration). The analyses by Sewell *et al*. (2015) also identified which signs in the *absence* of death are highly predictive of toxicity at the next higher concentration - some of these are more predictive than a single death at the lower concentration. Evident toxicity was subsequently defined as being reached if one or more animal displayed any of the following signs at least once from the day after the exposure: tremors, hypoactivity, >10% bodyweight loss and/or irregular respiration. This guidance has been incorporated in to TG433. However, the simplicity of this guidance has been questioned.

20. It is important to note that this is meant to act as a guide to help inform decisions on evident toxicity and no such guidance has been provided for other test guidelines using this endpoint. The experience of the study director and the presence of other signs should also be taken in account. It is the decision of the study director whether evident toxicity has been reached and there is the option to continue testing if this decision is unclear.

21. The dataset that this guidance was developed from has been extensively interrogated to look at the effect of combinations of signs, the duration of signs, and/or the number of animals displaying the sign(s) (Sewell *et al*., 2015). Since signs were most often observed in more than one animal and in combination with other signs this did not have a big impact on predictivity (some specific examples are provided in Sewell *et al*., 2015). Therefore the recommendations on recognising evident toxicity can be as simple as one sign observed at least once in one animal.

Effect of concentration ratio on the predictivity of signs

22. Additional analyses have been carried out to examine the effect of concentration ratio on the predictivity of the clinical signs used to provide guidance on the recognition of evident toxicity. The original analysis was carried out based on a minimum concentration change of 2-fold, but the fixed concentrations used in this test guideline (and in the accepted acute toxic class method, TG436) have varying ratios between them (e.g. 2-, 5- or 10-fold for dusts and mists). The majority of pairs of studies had a concentration ratio in the range of >2 to ≤5 (80%), with a substantial proportion of studies having a concentration ratio of >2 to ≤3 (39%). A smaller proportion of studies reported concentration ratios of >5 (20%) (Sewell *et al.* 2015). Additional analyses have been carried out to look at the predictivity of the most predictive clinical signs in different concentration ratio brackets >2, >5- and >10-fold (Appendix III). Predictivity increases with higher concentrations ratios, though confidence intervals widen (data for dusts and mists only).

23. Sewell *et al.* (2015) examined the effect of concentration ratios by comparing the concentration ratios for false positive (where the sign was observed but did not lead to toxicity at the next highest concentration) and true positive results i.e. to determine whether false positives were associated with lower concentration ratios. With the exception of faeces reduced, where false positive results were associated with a lower concentration ratio, there was no significant effect of concentration ratio on the occurrence of true/false positive results.

Signs observed in isolation versus multiple signs and/or multiple observations

24. There was concern that the guidance on evident toxicity could lead to over-classification by only requiring the observation of one sign in one animal. Though this situation is rare, the evidence shows that the four ‘evident toxicity’ signs are highly predictive even if only seen once in one animal, and even if no

other signs are observed in any other animal ('evident toxicity' signs or otherwise). For all evident toxicity signs that occurred in isolation (>10% bodyweight loss, hypoactivity and irregular respiration) and in only one animal this was still always associated with toxicity (death of 2 or more animals) at the next higher concentration. There were no exceptions. Tremors never occurred in isolation of other clinical signs.

Effect of severity of signs

25. There was concern that if a sign was seen only once in one animal, and the severity was mild, then this could result in over-classification. Severity was not consistently recorded in the dataset, only whether a sign was present or not**,** but the dataset should incorporate a range of severities. However, even considering that the dataset is likely to have included mild signs, if signs were only observed once in one animal this was still shown to be highly predictive of toxicity being observed at the next higher concentration. In the case that the decision on evident toxicity is unclear then there is the option to test at the higher concentration limit. The agreed wording for this option which has been incorporated into TG433 is *"In the event that a decision between Outcome B and C is being made in the absence of death, and the study director decides there is uncertainty as to whether evident toxicity has been observed (i.e. triggering Outcome B), there is the option to test at the next higher concentration limit".*

The choice of default sex

26. The original draft TG433 proposed the use of females only, unless males were indicated to be more sensitive. The sighting study was subsequently altered to test one male and one female to allow the identification of any large differences in sex sensitivity (as well as to inform a suitable starting concentration for the main study). It has been debated whether females really were the more sensitive sex as many considered males to be more sensitive for inhalation studies. The work by Price *et al.* (2011) indicated females were the more sensitive sex, but since information on the age and weight of animals was not included it is was questioned whether these factors could have influenced the results. There was no indication in the Sewell *et al.* (2015) dataset that either sex was more sensitive. Further analysis by the NC3Rs which examined the predictivity values of clinical signs for males versus females, including the occurrence of death, found no significant difference between sexes.

27. Since the general feeling was that males were likely to be more sensitive (potentially due to metabolic differences and lung capacity/minute volume), it was agreed that the guideline be changed to use males as default for this study. Only if the female appears to be much more sensitive than the male in the sighting study should this override the use of males for the main study.

Requirement for a sighting study

28. The requirement for a sighting study has also been questioned. It was agreed that though a sighting study is valuable to inform starting concentration it might not always be necessary to carry out, and therefore should not be compulsory. Therefore, if existing data are available to inform the starting concentration (and/or choice of sex), a sighting study is not necessarily be required. The wording in TG433 has been updated to reflect this.

Summary

29. All outstanding concerns regarding this test guideline have been fully addressed and agreed with the nominated inhalation toxicologists from the NL and US that raised the original concerns. A large and robust dataset has been used to develop guidance on the recognition of evident toxicity, an already accepted endpoint in existing TG, and extensive statistical evaluation and retrospective classifications support the adoption of this method for classification and labeling purposes. TG433 has been revised to incorporate the new data analysis and recommendations for evident toxicity. TG433 offers a significant

refinement to existing methods by using evident toxicity rather than death as an endpoint. It also uses fewer animals per study.

References

Price, C., N. Stallard, S. Creton, I. Indans, R. Guest, D. Griffiths, and P. Edwards. 2011. A statistical evaluation of the effects of gender differences in assessment of acute inhalation toxicity. *Human & experimental toxicology*. 30:217-238.

Sewell, F., I. Ragan, T. Marczylo, B. Anderson, A. Braun, W. Casey, N. Dennison, D. Griffiths, R. Guest, T. Holmes, T. van Huygevoort, I. Indans, T. Kenny, H. Kojima, K. Lee, P. Prieto, P. Smith, J. Smedley, W.S. Stokes, G. Wnorowski, and G. Horgan. 2015. A global initiative to refine acute inhalation studies through the use of 'evident toxicity' as an endpoint: towards adoption of the Fixed Concentration Procedure. *Regulatory toxicology and pharmacology*. 73. 770-779.

Stallard, N., C. Price, S. Creton, I. Indans, R. Guest, D. Griffiths, and P. Edwards. 2011. A new sighting study for the fixed concentration procedure to allow for gender differences. *Human & experimental toxicology*. 30:239-249.

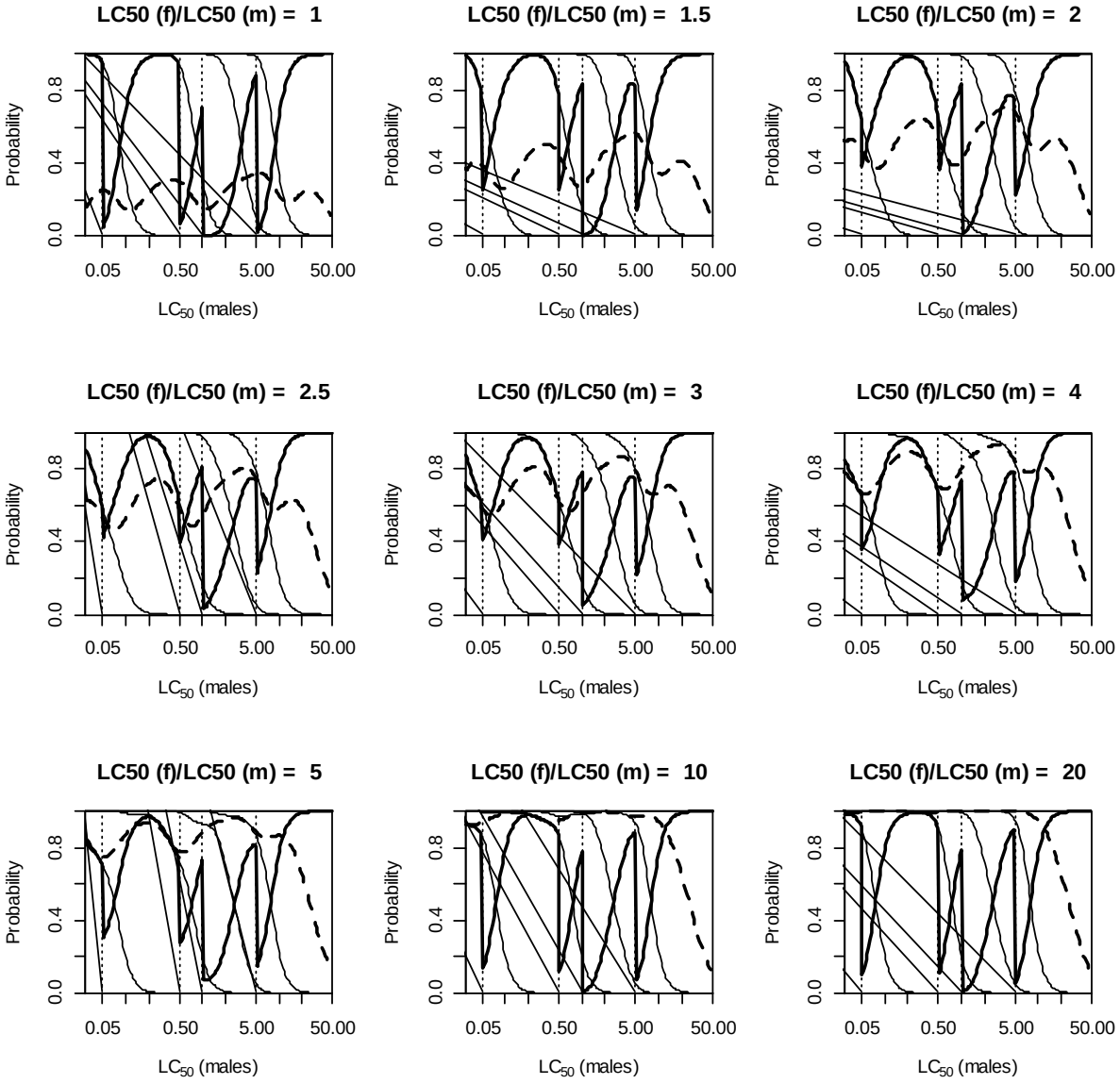
Stallard, N., A. Whitehead, and I. Indans. 2003. Statistical evaluation of the fixed concentration procedure for acute inhalation toxicity assessment. *Human & experimental toxicology*. 22:575-585.

van den Heuvel, M.J., D.G. Clark, R.J. Fielder, P.P. Koundakjian, G.J. Oliver, D. Pelling, N.J. Tomlinson, and A.P. Walker. 1990. The international validation of a fixed-dose procedure as an alternative to the classical LD50 test. *Food and chemical toxicology : an international journal published for the British Industrial Biological Research Association*. 28:469-482.

APPENDIX I (JANUARY 2016)

OECD TG 433 (FIXED CONCENTRATION PROCEDURE)

Classification probabilities for the fixed concentration procedure (FCP) with the new sighting study for dusts and mists with varying sex differences (males more sensitive), based on the current protocol. For more details on the plots and methods please refer to Stallard *et al.*, (2010).



These plots are similar to the plots in Stallard et al., (2010). The plots shows classification probabilities for the fixed concentration procedure (FCP) with the new sighting study for dusts and mists

with concentration-response curve slope of 4 and *R* (LC_{50}/TC_{50}) of 5 assuming a sighting study starting at 0.05mg/L. The solid heavy line gives the probability of the correct classification given the LC_{50}. The heavy dashed line gives the probability that the main study is conducted in the males rather than females.

The first plot corresponds to the case of no difference between the sexes. In this case the fact that the sighting study leads to a main study in females when no sex difference is observed means it is less likely the main study will be carried out in males (with the probability varying around 0.25). The other plots show what happens with increasingly large sex differences, with the males becoming more susceptible. In this case the LC_{50} on the *x*-axis is that for the males, as this is the true value on which classification should be based (since males are more sensitive), and the dashed line gives the probability that the main study is conducted in the males.

When the sex difference is small, there is quite a high change of erroneously testing in the females; for a ratio of LC_{50} values of 1.5 the probability is more than 0.5 in many cases. As the sex difference increases, the chance of seeing the sex difference in the sighting study and doing the main test in the males correctly also increases. For a ratio of LC_{50} values of 10 or more the probability of using males for the main test exceeds 0.9 except for the least toxic substances, when no effects are seen in either sex even at the highest test concentration, or extremely toxic substances, when deaths are seen in both sexes at the lowest test concentration.

The effect of sex differences with females more susceptible is less than that with males more susceptible since, as noted above, when there are no differences the main test is more likely to be conducted in the females. The effect of sex differences is also less when the concentration-response curve is steeper (and more when it is shallower). The work by Greiner et al. '*Report on Biostatistical Performance Assessment of Draft TG436 Acute Toxic Class Method for Acute Inhalation Toxicity'*, suggests that only about 1% of substances have a slope smaller than 4, showing that the plots above (with slope of 4) represent a worst-case scenario.

Reference

Stallard N, Price C, Creton S, Indans I, Guest R, Griffiths D, Edwards P (2010). A new sighting study for the fixed concentration procedure to allow for gender differences. *Human and Experimental Toxicology*. 30(3):239-49.

APPENDIX II (NOVEMBER 2016)

OECD TG 433 (FIXED CONCENTRATION PROCEDURE)

Retrospective classifications

Background: A number of publications have addressed comparability of the three acute inhalation methods - lethal concentration 50% (LC_{50}, OECD TG403), the acute toxic class (ATC) 'up-and-down method (OECD TG436) and the fixed concentration procedure (FCP) (OECD TG433). These used statistical simulations to compare the classifications made by each of the three methods and the likelihood of misclassification (under or over) (Price *et al*., 2011; Stallard *et al*., 2011; Stallard *et al*., 2003). In the absence of sex differences the statistical simulations showed that the three methods were comparable, but that all three methods had the potential to misclassify (Price *et al*., 2011). For all methods, classification is more accurate when the concentration-mortality curve is steep and performance is generally poorer for substances with shallower concentration-mortality curve, with classification being more variable.

For the FCP method, the statistical simulations assumed correct identification of evident toxicity to make the GHS classification predictions. Though evident toxicity is already an accepted endpoint in OECD TG420 (for the fixed dose procedure for acute oral toxicity studies), there were concerns that there may be variability and inconsistencies around the identification of evident toxicity in practice, which could potentially impact the classification. Through previous work we have provided guidance on the recognition of evident toxicity, so that this decision is more objective and consistent between laboratories (Sewell *et al*., 2015). Using this new, more objective definition of evident toxicity, we have used the same dataset to make retrospective classifications using the FCP method, to demonstrate how it can be used successfully in practice. We have compared the results of these retrospective classifications with retrospective classifications made using the two accepted methods (ATC and the LC_{50} methods).

Methods: The dataset collected by Sewell *et al*., (2015) was used to make retrospective classifications for dusts and mists (178 substances, and involving over 4000 animals). For each method, the classifications were established using the protocols and flow charts in their corresponding test guidelines, based on the starting concentration used in practice (Figures 1-3). For the FCP method classifications were made via females and males separately. For the LC_{50} method, rather than establish an LC_{50} value from the data a flowchart method was used based on whether more or less than 50% animals died at each concentration (as in Figure 1 in Price *et al*. 2011) (Figure 1). Where differences in sex sensitivity were identified classification was based on the most sensitive sex. Only 'valid' concentrations corresponding to within ±20% of the four fixed concentrations for dusts and mists in the ATC and FCP protocols (0.05, 0.5, 1 and 5 mg/L) were included, to comply with the guidelines. Retrospective classifications could only be made for substances where all the necessary and 'valid' concentrations were available. For example, for the FCP method, for a substance where testing started at 1mg/L and there was no death or evident toxicity in any animal, further testing would be required at 5mg/L. If this concentration was not been tested or the concentration fell outside of the ±20% then this substance could not be classified by this method.

Results and discussion: Retrospective classifications were made for 77 substances via the LC_{50} method, 57 substances via ATC, and 124 substances for FCP (101 substances using females only and 109

substances using males only). For FCP (for both males and females), classifications were generally able to be made using one or two concentrations requiring five to ten animals. For the ATC and LC_{50} methods classifications were generally made after two concentrations, requiring 12 and 20 animals respectively (Table 1).

There were 42 substances for which a retrospective classification was made via all four methods (LC50, ATC, FCP-females and FCP-males). For 35 of these (83.3%) all classifications were in agreement (Table 2). There were 7 substances for which there were disagreements between the classifications made by the different methods. Table 3 shows the data for these substances in more detail. If using the LC_{50} method as the 'reference' method (though it should be understood that there are limitations for this method and this also has the potential for misclassification), the ATC method under-classified by one class on three occasions. For the FCP method, when conducted in males only, there was one occasion of over-classification, and one of under-classification, both into the adjacent class. When the FCP was conducted in females, there was one occasion of over-classification into the adjacent more stringent class, and three occasions of under-classification, one of these by two classes (class 4 *vs.* class 2) (substance 5, Table 3). For this substance there were fewer female deaths at 1mg/L than at 0.5mg/L. Since 1mg/L was tested first, and the protocol allowed a classification to be made at that concentration, the results of the 0.5mg/L study were not taken in to account. For the 7 substances where there was a difference in the classifications made by the different methods there appeared to be a more sensitive sex for 6 of these, though these were not statistically significant (to be significant with a Fisher's exact test there would need to be no deaths for one sex vs. 4 or 5 deaths for the other, or 1 death in one sex vs. 5 in the other; any smaller difference is not enough evidence of differing sensitivity). For FCP, if the classification is made according to what appeared to be the more sensitive sex, there are fewer disagreements with the classifications from the LC_{50} method. Instead there are now only three occasions where classification made via FCP differs from LC_{50}, and these are all over-classifications into the adjacent more stringent class (it is important to note that the three occasions where the ATC method differed from the LC_{50} method were under-classifications into the less stringent adjacent class). This supports the conclusions from the previous statistical simulations that show that the FCP is comparable to the existing methods if sex differences are taken in to account. If the FCP classifications were based on females as the default sex there are 5 occasions where classifications differ from those made by LC50 and 4 occasions where the classifications differ from those made by the ATC method.

Often it was not possible to make a classification via all methods (e.g. due to a missing concentrations), and there are more examples of where comparisons could be made where classifications could be made by two of the methods. Table 4 compares the classifications made by combinations of two methods, showing that there was over 90% agreement for all combinations. Tables 5-9 show the classifications made for each combination in more detail, to show how these differed.

Discussion/conclusions: There is a strong agreement between the classifications made by each of the three methods. It is encouraging that there is strong agreement between the classifications made by the FCP and the two accepted methods, irrespective of the sex used by FCP. This shows that our definition of evident toxicity can be used in practice. The FCP method allows classification to be made using fewer studies and fewer animals. It also offers animal welfare benefits through the avoidance of death as an endpoint.

Tables and Figures

Table 1: Number of studies required to allow classification, and the number of substances classified.

No. concentrations required to make a classification	Number of substances			
	FCP (females)	FCP (males)	ATC (TG436)	LC50 (TG403)
1 study	54	64	18	32
2 studies	46	41	37	41
3 studies	1	3	2	3
4 studies	0	1	0	1
Total no. substances classified	**101**	**109**	**57**	**77**

Table 2: Classifications made by all three methods, showing the number of substances classified in to each class and the number of substances where there was a disagreement between the three methods (which is expanded on in Table 3).

Classification	**No. substances**
Class 1	1
Class 2	11
Class 3	3
Class 4	14
Class 5	6
Disagreements	7

Table 3: Retrospective classifications made for substances where there were disagreements in the classifications made via the different methods, showing concentrations tested and the number of deaths or animals with evident toxicity at each concentration. Retrospective classifications were based on the order the studies were carried out.

Substance	**Concentrations tested**		**No. deaths**		**No. evident toxicity**		**Classification**			
			F	M	F	M	LC_{50}	ATC	FCP(F)	FCP(M)
1	START	0.5mg/L	0	0	0	0	3	4	3	4
		1 mg/L	4	1	1	0				
2	START	5 mg/L	2	1	3	4	5	5	4.	5*
		1 mg/L	0	0	4	4				
3	START	5 mg/L	0	2	5	3	5	5	5*	4
		1 mg/L – males	-	0	-	0				
4	START	5 mg/L	0	3	5	2	4	5	5*	4
		1 mg/L – males	-	0	-	5				
5	START	1 mg/L	1	4	2	0	2	2	4*.	2
		0.5 mg/L	3	4	0	0				
		0.05 mg/L	0	0	0	0				
6	START	1 mg/L	0	0	0	0	5	5	4.	5
		5 mg/L	2	0	3	5				
7	START	5 mg/L	5	5	0	0	3	4	$4^{\#}$	3
		1 mg/L	1	3	4	2				
		0.5 mg/L - males	-	0	-	0				

* Classification made after first concentration tested. $^{\#}$ Classification made after second concentration tested.

Table 4: Comparisons of the classifications made via the different methods (FCP in females (FCP-F), FCP in males (FCP-M), ATC and LC50), showing the number of substances classified by both methods and the number of studies in agreement.

Comparison			**No. classified**	**No. studies in agreement**	**% agreement**
ATC	*vs.*	FCP-F	46	42	91.3%
LC_{50}	*vs.*	FCP-F	43	40	93.0%
LC_{50}	*vs.*	FCP-M	44	41	93.2%
ATC	*vs.*	FCP-M	51	48	94.1%
LC_{50}	*vs.*	ATC	46	44	95.7%

Table 5: Comparison of classifications made by FCP in females (FCP-F) and by the ATC method. Classifications were made for 46 substances, 42 (91.3%) of which were in agreement for both methods.

		ATC				
	Class	**Class 1**	**Class 2**	**Class 3**	**Class 4**	**Class 5**
FCP-F	**Class 1**	1	0	0	0	0
	Class 2	0	11	0	0	0
	Class 3	0	0	3	1	0
	Class 4	0	1	0	17	2
	Class 5	0	0	0	0	10

Table 6: Comparison of classifications made by FCP in females (FCP-F) and by the LC_{50} method. Classifications were made for 43 substances, 40 (93.0%) of which were in agreement for both methods.

		LC_{50}				
	Class	**Class 1**	**Class 2**	**Class 3**	**Class 4**	**Class 5**
FCP-F	**Class 1**	1	0	0	0	0
	Class 2	0	11	0	0	0
	Class 3	0	0	5	0	0
	Class 4	0	0	1	15	2
	Class 5	0	0	0	0	8

Table 7: Comparison of classifications made by FCP in males (FCP-M) and by the LC_{50} method. Classifications were made for 44 substances, 41 (93.2%) of which were in agreement for both methods.

		LC_{50}				
	Class	**Class 1**	**Class 2**	**Class 3**	**Class 4**	**Class 5**
FCP-M	**Class 1**	1	0	0	0	0
	Class 2	0	11	0	0	0
	Class 3	0	0	4	0	0
	Class 4	0	0	2	14	1
	Class 5	0	0	0	0	11

Table 8: Comparison of classifications made by FCP in males (FCP-M) and by the ATC method. Classifications were made for 51 substances, 48 (94.1%) of which were in agreement for both methods.

		ATC				
	Class	**Class 1**	**Class 2**	**Class 3**	**Class 4**	**Class 5**
FCP-M	**Class 1**	1	0	0	0	0
	Class 2	0	12	0	0	0
	Class 3	0	0	3	1	0
	Class 4	0	0	0	16	2
	Class 5	0	0	0	0	16

Table 9: Comparison of classifications made by the LC_{50} and ATC methods. Classifications were made for 46 substances, 44 (95.7%) of which were in agreement for both methods.

		LC_{50}				
	Class	**Class 1**	**Class 2**	**Class 3**	**Class 4**	**Class 5**
ATC	**Class 1**	1	0	0	0	0
	Class 2	0	11	0	0	0
	Class 3	0	0	3	0	0
	Class 4	0	0	2	16	0
	Class 5	0	0	0	0	13

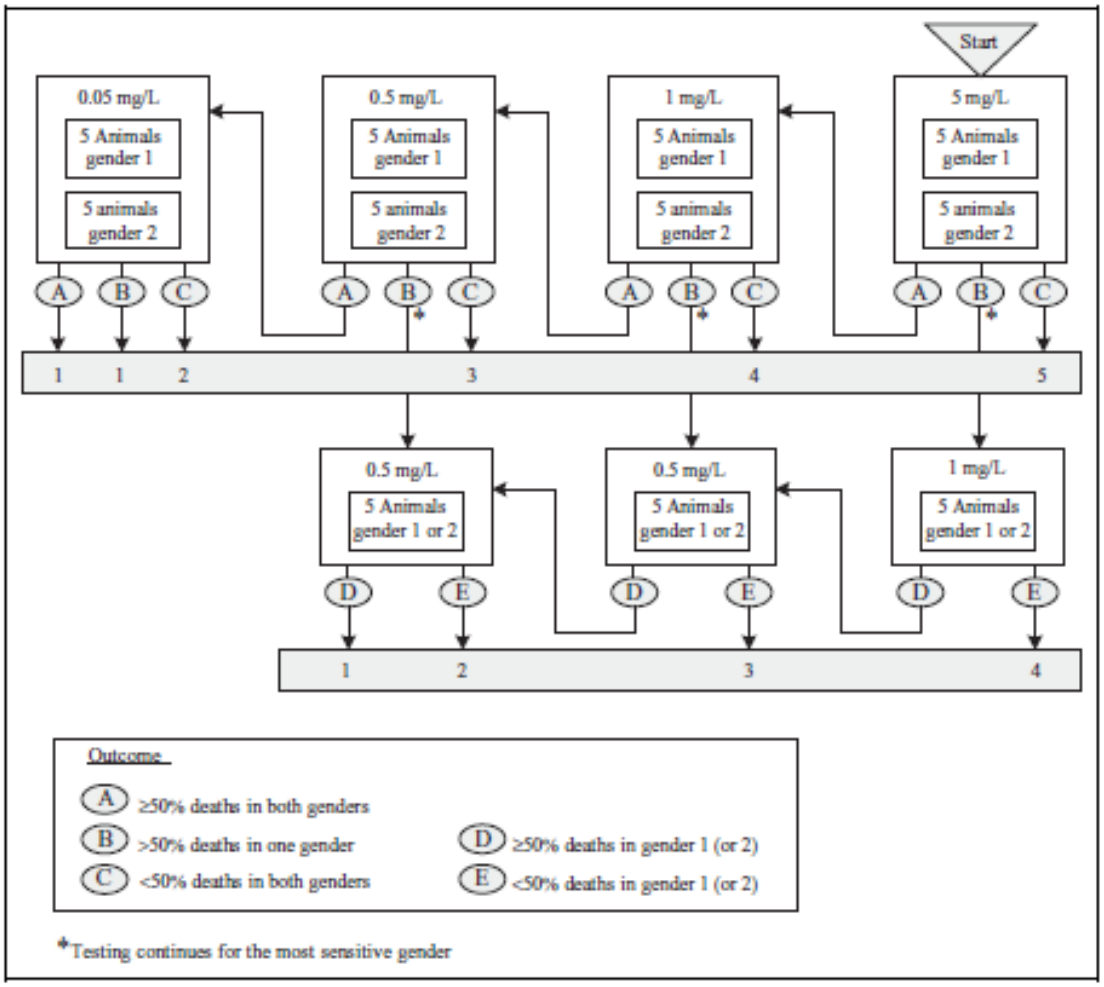


Figure 1: LC50 test for dusts and mists starting at 5mg/L (Price *et al*., 2011).

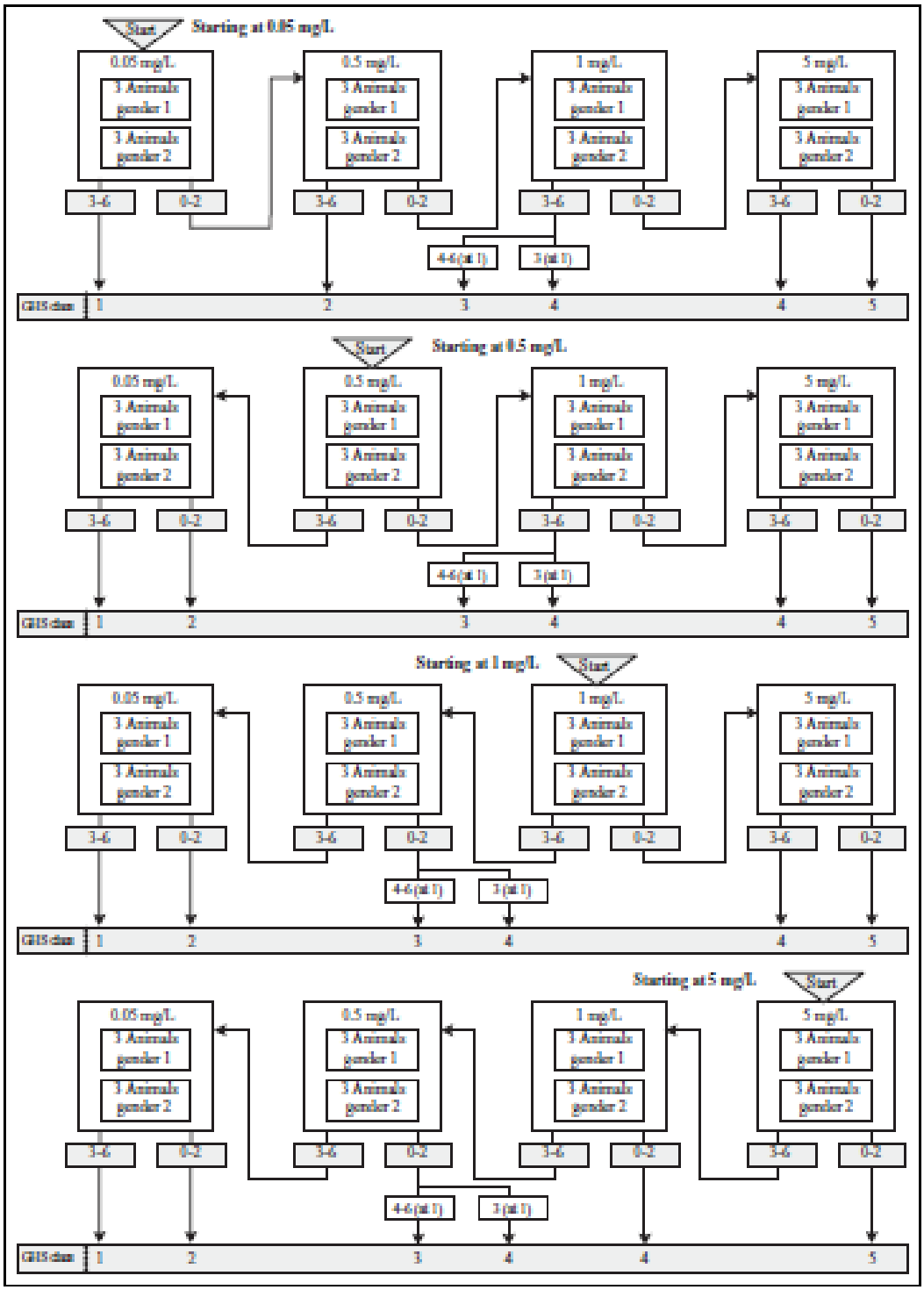


Figure 2: ATC study for dusts and mists (Price *et al.*, 2011).

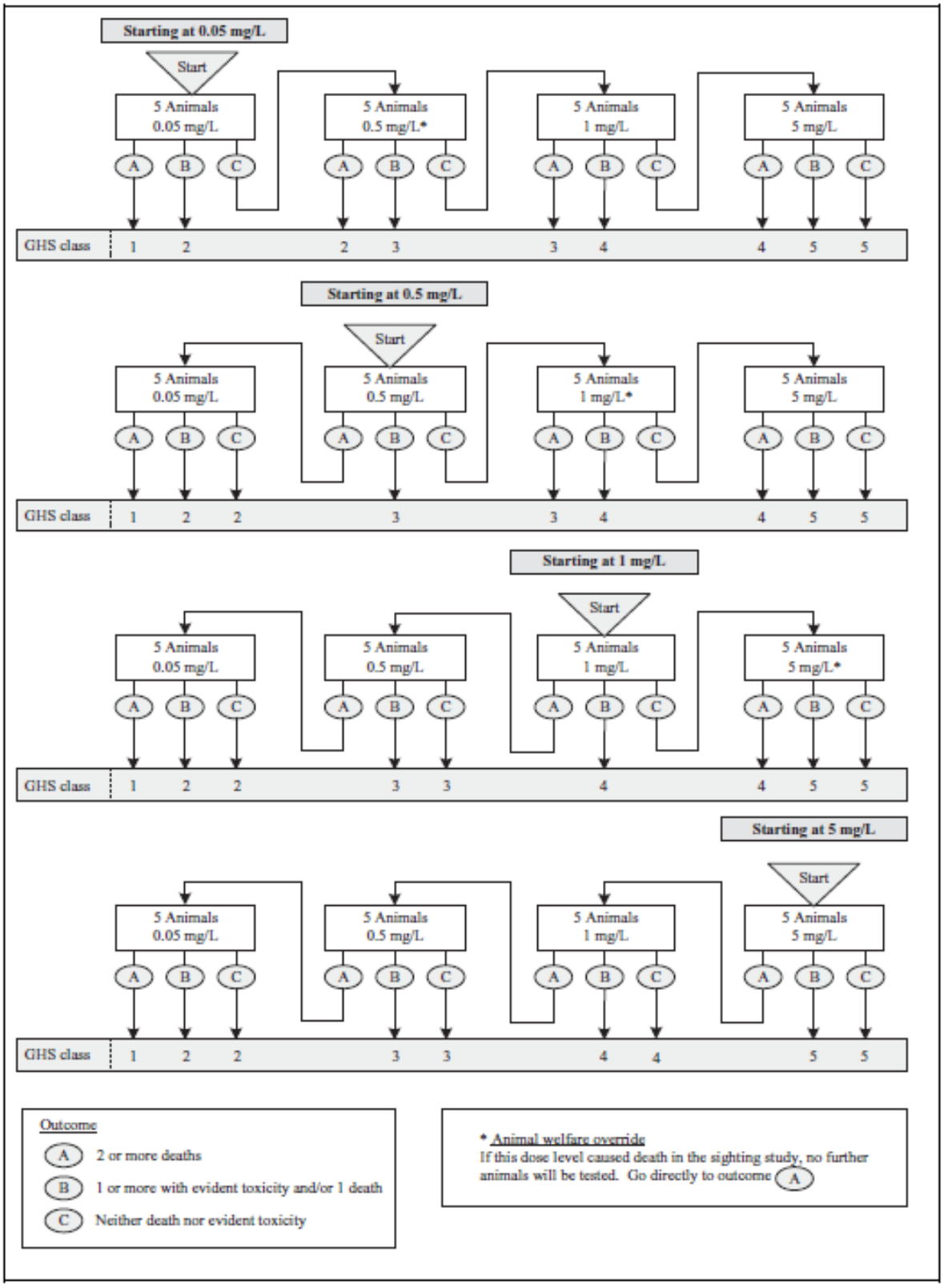


Figure 3: FCP main study protocol for dusts and mists (Price *et al*., 2011).

References

Price, C., N. Stallard, S. Creton, I. Indans, R. Guest, D. Griffiths, and P. Edwards. 2011. A statistical evaluation of the effects of gender differences in assessment of acute inhalation toxicity. *Human & experimental toxicology*. 30:217-238.

Sewell, F., I. Ragan, T. Marczylo, B. Anderson, A. Braun, W. Casey, N. Dennison, D. Griffiths, R. Guest, T. Holmes, T. van Huygevoort, I. Indans, T. Kenny, H. Kojima, K. Lee, P. Prieto, P. Smith, J. Smedley, W.S. Stokes, G. Wnorowski, and G. Horgan. 2015. A global initiative to refine acute inhalation studies through the use of 'evident toxicity' as an endpoint: Towards adoption of the fixed concentration procedure. *Regulatory toxicology and pharmacology*. 73:770-779.

Stallard, N., C. Price, S. Creton, I. Indans, R. Guest, D. Griffiths, and P. Edwards. 2011. A new sighting study for the fixed concentration procedure to allow for gender differences. *Human & experimental toxicology*. 30:239-249.

Stallard, N., A. Whitehead, and I. Indans. 2003. Statistical evaluation of the fixed concentration procedure for acute inhalation toxicity assessment. *Human & experimental toxicology*. 22:575-585.

APPENDIX III (JANUARY 2017)

OECD TG 433 (FIXED CONCENTRATION PROCEDURE)

Additional analyses

Background: Following the teleconference in November 2016 and subsequent feedback, we agreed to carry out additional analyses to clarify some of the remaining issues regarding acceptance of this guideline, mainly relating to:

- The effect of concentration ratio on the ability of clinical signs observed at the lower concentration to predict toxicity (death of two or more animals) at the higher concentration.
- The effect of the severity of clinical signs observed at the lower concentration on predictivity of toxicity at the next highest concentration, as well as the duration or number of animals experiencing the sign.
- A comparison of combinations of signs that indicate evident toxicity has occurred, versus signs that are observed in isolation.

Concentration ratios:

Additional analyses have been carried out to examine the effect of concentration ratio on the predictivity of the clinical signs used to provide guidance on the recognition of evident toxicity. The fixed concentrations used in this test guideline (and in the accepted acute toxic class method, TG436) have varying ratios between them (e.g. 2-, 5- or 10-fold for dusts and mists; Table 1). For classification of dusts and mists, the lowest concentration change occurs between GHS categories 2 and 3, with a 2-fold difference. It was suggested that it would be valuable to look at the effect the fold change has on the predictivity of the clinical signs observed at the lower concentration, as presumably the greater the concentration change the more likely it is that toxicity will occur at the higher concentration.

Table 1: GHS classifications for LC_{50} by inhalation.

GHS category	Vapours (mg/L)	Dusts and mists (mg/L)	Gases (ppm)
1	≤0.5	≤0.05	≤100
2	>0.5 and ≤2	>0.05 and ≤0.5	>100 and ≤500
3	>2 and ≤10	>0.5 and ≤1	>500 and ≤2,500
4	>10 and ≤20	>1 and ≤5	>2.500 and ≤20,000
5 (unclassified)	20	5	>20,000

This has already been addressed to some extent in the publication by Sewell *et al.* (2015). In the dataset upon which the guidance on evident toxicity has been developed a minimum concentration change of 2-fold was used to reflect the concentration changes in the test guidelines (TG436 and TG433). Indeed, the majority of studies had a close to 2-fold concentration change, so the data is in fact skewed towards a more conservative, smaller, concentration change. The majority of pairs of studies had a concentration ratio in the range of >2 to ≤5 (80%), with a substantial proportion of studies having a concentration ratio of >2 to ≤3 (39%). A smaller proportion of studies reported concentration ratios of 5 or more (20%) (Sewell *et al.* 2015).

Table 11 in Sewell *et al.* (2015) (copied below), looked at the clinical signs used to provide guidance on assessment of evident toxicity (hypoactivity, tremors, bodyweight loss, and irregular respiration) as well as some other highly predictive signs (body staining, ano-genital staining, faeces reduced, naso-ocular discharge, noisy respiration and hunched posture) to examine the concentration ratios for false positive (where the sign was observed but did not lead to toxicity at the next highest concentration) and true positive results i.e. to determine whether false positives were associated with lower concentration ratios. With the exception of faeces reduced, where false positive results were associated with a lower concentration ratio, there was no significant effect of concentration ratio on the occurrence of true/false positive results.

Table 11
Concentration ratios (the ratio between the lower and higher dose) and PPVs for commonly observed clinical signs. For each of the signs shown, average concentration ratios in those studies giving rise to false positive prediction of toxicity were compared with the average concentration ratios in those studies giving rise to true prediction of toxicity. The p-value is calculated from a t-test of whether the mean concentration ratio differs between false and true positives.

Clinical sign	Code	PPV (95% CI)		Concentration ratios		p-value
				False positive (sign present but toxicity does not occur)	True positive (sign present & toxicity occurs)	
Hypoactivity	L	100.0	(92.4–100.0)	–	4.09	–
Tremors	T	100.0	(68.8–100.0)	–	4.56	–
Bodyweight loss	BW	94.0	(84.6–98.4)	4.97	3.57	0.610
Irregular respiration	IR	89.0	(80.9–94.5)	4.30	5.14	0.296
Body staining	ST	88.5	(71.8–97.0)	2.06	3.61	0.001
Ano-genital staining	UGS	86.4	(67.3–96.4)	4.71	3.50	0.671
Faeces reduced	FR	85.3	(70.4–94.4)	2.07	3.89	<0.001
Naso-ocular discharge	ND	85.0	(71.4–93.7)	3.40	4.27	0.184
Noisy respiration	RN	81.2	(71.9–88.4)	3.62	3.90	0.597
Hunched posture	H	78.8	(66.3–88.3)	2.55	3.57	0.011

We have carried out additional analyses for the dusts and mists included in our dataset to examine the positive predictive values (PPV) for the highly predictive signs at concentration fold changes of ≥2, ≥5 and ≥10 to reflect the concentration fold changes used for GHS classification of dusts and mists. Larger concentration ratios do show some improvement in predictivity, but the data pool for these studies is smaller so this is associated with wider confidence intervals, and lower sensitivity values. Irregular respiration is particularly improved with increasing concentration ratio, with consistent confidence intervals. However, this sign is still highly predictive for just a 2-fold concentration change.

Table 3: PPV (95% confidence interval) for highly predictive signs with 2, 5 or 10-fold concentration change between the lower and higher concentration.

Clinical sign	**≥2-fold** (95% CI)		**≥5-fold** (95% CI)		**≥10-fold** (95% CI)	
Tremors	100.0	(68.8 - 100.0)	100.0	(5.0 - 100.0)	100.0	(5.0 - 100.0)
Hypoactivity	100.0	(92.0 - 100.0)	100.0	(47.3 - 100.0)	100.0	(47.3 - 100.0)
>10% bodyweight loss	91.7	(79.0 - 97.8)	85.7	(47.0 - 99.3)	100.0	(36.8 - 100.0)
Irregular respiration	89.0	(80.9 - 94.5)	95.8	(81.2 - 99.8)	100.0	(86.1 - 100.0)
Body staining	88.5	(71.8 - 97.0)	100.0	(60.7 - 100.0)	100.0	(22.4 - 100.0)
Ano-genital staining	86.4	(67.3 - 96.4)	0.0	(0.0 - 95.0)	100.0	(5.0 - 100.0)
Faeces reduced	85.3	(70.4 - 94.4)	100.0	(47.3 - 100.0)	100.0	(47.3 - 100.0)
Naso-ocular discharge	84.2	(70.1 - 93.3)	100.0	(74.1 - 100.0)	100.0	(65.2 - 100.0)
Noisy respiration	80.5	(70.9 - 88.0)	94.1	(74.3 - 99.7)	100.0	(68.8 - 100.0)
Hunched posture	78.0	(65.0 – 87.8)	87.5	(64.5 - 97.8)	100.0	(54.9 - 100.0)
Gasping	76.5	(52.5 – 92.0)	100.0	(22.4 - 100.0)	100.0	(22.4 - 100.0)

Scale of severity:

It was questioned whether information on the severity of observed signs, duration and/or number of animals exhibiting the sign could be included as part of the guidance on the recognition of toxicity. Irregular respiration, one of the most highly predictive signs that indicates evident toxicity, was mentioned as a particular concern. As part of the guidance development process, extensive sub-analyses for duration of

signs, number of animals experiencing the sign and, where possible, grading of signs was carried out and was found not have a great effect on predictivity. Though increasing the numbers of animals exhibiting a sign or the number of observations in an individual animal did increase the predictivity of some clinical signs to some extent, it also widened the confidence intervals, due to the smaller data pool. For these reasons the guidance on the recognition of evident toxicity was simply defined as the observation of any of the most highly predictive signs of evident toxicity (>10% bodyweight loss, tremors, irregular respiration and hypo activity) at least once, in at least one animal after the day of exposure. Some of the sub-analyses are included in the publication by Sewell *et al.* (2015). Table 7 in the publication (copied below) looked at the effect of increasing the number of animals with irregular respiration, showing that it did not impact predictivity.

Table 7
The effect of changing the number of animals experiencing clinical sign irregular respiration (IR) on predictivity (PPV).

No. animals with sign IR	PPV (95% CI)		Sensitivity (95% CI)		Specificity (95% CI)		No. studies
1	89.0	(80.9–94.5)	35.3	(29.0–42.0)	85.2	(74.7–92.5)	82
2	87.5	(78.3–93.7)	30.4	(24.5–36.9)	85.2	(74.7–92.5)	72
3	86.4	(76.5–93.1)	27.5	(21.8–33.9)	85.2	(74.7–92.5)	66
4	84.5	(73.5–92.1)	23.7	(18.3–29.8)	85.2	(74.7–92.5)	58
5	85.1	(72.8–93.2)	19.3	14.4–25.1)	88.5	(78.7–94.8)	47

Though we note this is somewhat different from the gradings of other clinical signs, Table 9 in the Sewell *et al.* (2015) publication (also copied below), Iooked at the different levels of bodyweight loss reported in the retrospective dataset and showed that this also did not have a great impact on predictivity. A bodyweight loss of >10% compared to the day of exposure (day 0), was shown to be highly predictive of toxicity at the next highest concentration, from the day after exposure (day 1 onwards).

Table 9
Effect of sub-categories of bodyweight loss on PPV. Sign observed at least once in at least one animal.

Clinical sign	Code	PPV (95% CI)		Sensitivity (95% CI)		Specificity (95% CI)		No. studies
Bodyweight loss (unspecified)	BW (unspecified)	100.0	(77.9–100.0)	5.8	(3.2–9.6)	100.0	(95.2–100.0)	12
Bodyweight loss (mild)[a]	BW (mild)	100.0	(36.8–100.0)	1.4	(0.4–3.9)	100.0	(3.9–95.2)	3
Bodyweight loss (moderate)[b]	BW (moderate)	94.4	(82.9–99.0)	16.4	(11.9–21.9)	96.7	(21.9–89.6)	36
Bodyweight loss (substantial)[c]	BW (substantial)	100.0	(22.4–100.0)	1.0	(0.2–3.1)	100.0	(3.1–95.2)	2
Thin	Thin	50.0	(2.5–97.5)	0.5	(0.1–2.3)	98.4	(2.3–92.2)	2

[a] Reduced weight gain.
[b] Weight loss 10–20% compared to pre-dosing weight on day 0.
[c] Weight loss >20% compared to pre-dosing weight on day 0.

It has been noted that the pre-study acclimatisation procedure has an influence on the body weight loss recorded the day after exposure. Since the dataset came from a number of different laboratories this should incorporate a range of different acclimatisation procedures. It is, however, the judgement of the study director as to whether they feel that the acclimatisation procedure has influenced the bodyweight loss observed and therefore bodyweight loss could instead be used as a marker from day 2 . The guidance on interpreting evident toxicity is not intended to be definitive, but simply a guide to inform decision-making.

Since the historical data used for the analysis was generated in a number of different laboratories, information on grading was not always available for the majority of signs, and it was therefore not possible to incorporate a consistent grading system. For the purposes of the analyses the lowest common denominator was used (i.e. if the sign was observed enough to have been recorded, it was included in the analyses). Since the data came from a large number of studies, involving a number of different laboratories (and presumably an even greater number of personnel/technical staff), a significant amount of subjectivity - as well as varying severities - will already be accounted for. For the evident toxicity signs included in the guidance (hypoactivity, tremors, bodyweight loss, and irregular respiration), irrespective of the severity, if these signs were recorded they were highly predictive of toxicity at the next highest concentration. Furthermore, the retrospective analyses carried out showed that there is still strong agreement in the classifications made using all three methods.

It has also been noted that, since all the studies in the retrospective dataset were carried out with mortality as the endpoint, and clinical observations did not directly influence the study outcome, it is possible that changing the study endpoint to evident toxicity could alter the way that this data is recorded. Again, this reflects the recurring issue of subjectivity surrounding the use of clinical signs. It is intended that

the signs are recorded in exactly the same way as with the accepted guidelines (TG403 and TG436). Though it has been suggested that this change in focus may influence recording and decision making to some extent and for some individuals, we do not believe this will be a significant problem as the retrospective dataset already incorporates a large level of subjectivity. If those involved in the conduct or analysis of the study are unsure, the test guideline will include the option to test at a higher concentration if deemed necessary, but we hope that this would be a rare occurrence, particularly as confidence grows in using this method.

Signs observed in isolation versus combinations of signs:

It has been questioned whether the guidance on recognition of evident toxicity could include combinations of signs that indicate evident toxicity has occurred, rather than individual signs. As part of the guidance development process, this was considered and extensive sub-analyses carried out to examine this. Combinations of signs were not shown to significant increase the predictivity.

There was concern that a situation where only one of the evident toxicity signs seen once and in only one of the five animals could potentially result in over-classification, particularly if the sign was mild in severity. Irregular respiration was specified as of particular concern for this situation. In the historical dataset analysed it was rare that signs occurred in isolation, only once, and only in one animal. The evident toxicity signs were usually observed multiple times in multiple animals, which is why analyses that examined the impact of increasing the number of observations or animals exhibiting a sign did not lead to a marked improvement in predictivity. Similarly, as it was rare to observe signs in isolation, combinations of signs also did not have a great impact on predictivity. These are further reasons why the definition of evident toxicity can be simply defined as the observation of one sign at least once in at least one animal from the day after exposure.

We have re-examined the dataset to determine how often evident toxicity signs occurred in isolation (no other signs observed), and only once in one of the group of five animals, and how this would potentially influence the classifications made.

Only 12 signs occurred in isolation in a single animal, some of these very infrequently (Table 4). Irregular respiration was the only sign that was frequently observed in isolation, 42% of occurrences were not associated with any other sign. However, it was rare that the sign was seen in only one of the five animals. Of the 268 pairs of studies included in the main prediction analysis (pairs of studies with >2-fold concentration change, with no deaths at the lower concentration), there were only 10 studies in which irregular respiration was observed in only one of the 5 animals in the group, and in only 5 of these studies were no other clinical signs recorded in any other animal. For these 5 studies there was always toxicity at the next highest concentration (death of two or more animals) (Table 5). These were all female studies, but males were also tested for 4 of these studies. The results of all the studies carried out for these five substances are summarised in Table 5, showing the number of deaths and/or animals displaying evident toxicity for each study, and additional information on the clinical signs observed.

This example shows that the observation of just a single ‘evident toxicity’ sign in the group of 5 animals is unlikely to result in over-classification. For irregular respiration, the sign most likely to be observed in isolation, all 5 substances would have been correctly classified. However, if a single ‘evident toxicity’ sign is observed in a single animal and the study director is uncertain as to whether evident toxicity has been observed, there is the option to continue the study to observe for development of further signs and/or to test at the next highest concentration if required.

Table 4: Number of animals displaying a clinical sign in isolation, and the total number of animals displaying the sign.

Clinical sign	**No. animals displaying sign ONLY**	**% total no. animals displaying the sign**	**Total no. animals displaying the sign**
Irregular respiration	137	42%	325
Body staining	27	27%	99
Hypoactivity	12	16%	77
Laboured respiration	12	16%	77
Faeces reduced	13	12%	107
Hunched posture	18	8%	227
Ano-genital staining	4	8%	51
Naso-ocular discharge	6	7%	89
Congested respiration	4	5%	87
Facial staining	3	5%	65
>10% bodyweight loss	2	2%	93
Noisy respiration	1	0.4%	267

Retrospective classifications:

Prior to the previous teleconference a summary of the results of the retrospective classifications was shared, then presented during the teleconference. There was a strong agreement between the classifications made using each of the three methods. For FCP (for both males and females), classifications were generally able to be made using one or two concentrations requiring five to ten animals, whereas the other two methods usually required a greater number of studies (and a greater number of animals tested) before a classification could be made.

Since the original classifications made following each study were not available for many of the substances, retrospective classifications were made by all three methods for all substances. Additional information on the methodology used for this was requested and the classification rules for each method have been provided in an Excel file to show the decision making process. For each method, the classifications were established using the protocols and flow charts in their corresponding test guidelines, based on the starting concentration used in practice (Figures 1-3).

For the LC_{50} method, rather than establish an LC_{50} value from the data, a flowchart method was used based on whether more or less than 50% of animals died at each concentration (as in Figure 1 of Price *et al*., 2011) (copied below). This method will have limitations as the retrospective classifications will be based on only the fixed concentrations (i.e. 0.05, 0.5, 1 and 1 mg/L for dusts and mists), and will omit data from any other concentrations tested (though it was rare that other concentrations were used, 2 mg/L was used in approximately 15 studies). It should be noted that only 'valid' concentrations (corresponding to within ±20% of the four fixed concentrations of 0.05, 0.5, 1 and 5 mg/L for dusts and mists in the ATC and FCP protocols) were included, to comply with the guidelines. Retrospective classifications could therefore only be made for substances where all the 'valid' concentrations were reported. For example, using the FCP method, for a substance where testing started at 1mg/L and no death or evident toxicity occurred in any animal, further testing would be required at 5mg/L. If this concentration (±20%) had not been tested then a substance could not be classified by this method.

Information on original classifications, made via the LC_{50} method, was available for 110 substances. We have carried out additional analyses to compare these original and our retrospective classifications. Due

to availability of all the required data, we were only able to make retrospective classifications for 25 of these substances, but these show strong concordance (23/25 classifications agreed). It should, however, be noted that for many of these original classifications the decision was not always distinct, with many of the classifications given as class 3/4 (25 substances) or class 2/3 (1 substance), presumably because the LC_{50} value fell near the class border. We were unable to make retrospective classifications for such substances using the LC_{50} method, but were able to make retrospective classifications via the FCP method for 22 of the 25 substances that were classified as class 3/4. All except of two substances gave retrospective classifications of 3, the more conservative of 3/4. The other two substances gave a classification of 2. Both these substances had also been tested at 2 mg/L (which could not be included in the retrospective classifications as this is not within ±20% of the fixed concentrations), which resulted in death of almost all animals (9/10 deaths for one substance, and 10/10 for the other). At 0.5 mg/L there were 2 deaths in the female groups for both substances and one male death for one substance and evident toxicity in all other animals. No deaths and no evident toxicity occurred at 0.05 mg/L.

The strong agreement between the retrospective classifications made by each of the three methods, and the strong concordance with the original classifications provides further evidence (in addition to the statistical simulations in the publications by Stallard *et al*. 2011 and Price *et al.*, 2011) that the FCP (TG433) method can be reliably be used to make classifications, with minimal potential for over or under classification.

Table 5: Studies where irregular respiration was observed only once in one animal at the lower concentration in females, with no other signs.

	Conc. tested	Female observations			Male observations		
		No. deaths	No. evident toxicity	Additional notes/observations	No. deaths	No. evident toxicity	Additional notes/observations
1	0.05 mg/L	0	1	IR in 1 animal, no other signs in any other animal.	0	4	IR in 4 animals, no signs in the other animal.
	0.5 mg/L	5	-	Multiple signs preceding death (including IR, L, RG, ND)	3	2	Multiple signs preceding death, and in the remaining animals (including IR, L, RG, ND, FAS, BW)
	2 mg/L	5	-	Multiple signs preceding death (including IR, L, RG, ND)	5	0	Multiple signs preceding death (including IR, L, RG, ND, DA, PO)
2	0.06 mg/L	0	1	IR in 1 animal, no other signs in any other animal.	0	5	IR in all animals for more than two days, also BW in one animal.
	0.5 mg/L	2	3	Multiple signs in remaining animals (including IR, RG, BW)	3	2	Multiple clinical signs preceding death and in remaining animals (including IR, RG, UGS, BW, DA)
	2 mg/L	4	1	Multiple signs preceding death and in the remaining animal (including IR, L, ND, RL, FAS, H, UGS FR, BW)	5	-	Multiple clinical signs preceding death (including IR, L, T, ND, RL, FAS, H, UGS, FR, BW)
3	0.5 mg/L	0	1	IR in 1 animal, no other signs in any other animal.	0	4	IR in 4 animals, additional signs in 2 animals (including FAS, FR), no signs in 1 animal
	2 mg/L	2	3	Deaths on day of exposure, multiple signs in remaining animals (including L, H, FAS)	2	3	Deaths on day of exposure, multiple signs in remaining animals (including L, H, FAS)
4	0.05 mg/L	0	1	IR in 1 animal, no other signs in any other animal.	0	2	IR in 2 animals, no other signs in other animals.
	0.2 mg/L	5	-	Multiple signs preceding death (including L, IR, ND, FR, FAS, H, BW)	5	-	Multiple signs preceding death (including L, IR, ND, FR, FAS, H, BW)
	2 mg/L	5	-	Multiple signs preceding death (including IR, L, H FAS, UGS, FR, H, RL, BW)	5	-	Multiple signs preceding death (including L, IR, H, FAS, UGS, FR, BW)
	5 mg/L	5	-	Multiple signs preceding death (including	5	-	Multiple signs preceding death (including

				IR, L, UGS, CTT, BW)			IR, ND, L, H, UGS, CTT, FR, BW)
5	0.06 mg/L	0	1	IR in 1 animal, no other signs in any other animal.	n/a	n/a	Males not tested
	0.5 mg/L	2	3	Multiple clinical signs preceding death and in remaining animals (including IR, UGS, FR, FAS, BW)	0	5	Multiple clinical signs in all animals (including IR, UGS, FR, FAS, BW)
	2 mg/L	5	-	All died, within 1h exposure	5	0	All died, within 1h exposure

Clinical signs: IR, irregular respiration; L, hypoactivity; RG, gasping; ND, naso-ocular staining; BW, bodyweight loss (>10%), FAS, facial staining; DA, distended abdomen; PO, prone posture; H, hunched posture; FR; faeces reduced;RL, laboured respiration; T, tremors; CTT, cold to touch

Summary

A large evidence-base of clinical signs was collected and extensively analysed in order to develop guidance on the recognition of evident toxicity to support the Fixed Concentration Procedure (OECD TG433) for acute inhalation studies (Sewell *et al.*, 2015). Evident toxicity has been reached if any of the four 'evident toxicity signs' (i.e. bodyweight loss (>10%), tremors, hypoactivity or irregular respiration) are observed at least once in at least one animal from the day after exposure - it is highly likely that toxicity (death of at least 2 animals) will occur at the next highest concentration. Though extensive sub-analyses have been carried out to develop this guidance, the simplicity of the recommendations has been questioned. The recommendations apply to a minimum number of observations (observed at least once) and the minimum concentration change (at least 2-fold), but predictivity may increase with larger concentration change. Increasing the number of animals experiencing the sign did not have a great impact on predictivity as signs usually occurred in multiple animals. Similarly, combinations of signs did not have a great impact on predictivity, as signs did not usually occur in isolation. It was not possible to look at the severity of signs in great detail as this information was not always available, but the dataset used to make these recommendations will incorporate a large level of subjectivity and grading of clinical signs – if it was observed enough to be recorded the 'evident toxicity signs' are predictive of toxicity at the next highest concentration.

The ability of the 'evident toxicity signs' to predict toxicity when only observed once in one animal with no other accompanying signs was of particular concern, particularly for irregular respiration. However, this situation happened rarely in the dataset (5/268 (<2%) studies) and in all 5 cases this was predictive of toxicity at the next highest concentration.

The recommendations for the recognition of evident toxicity are not intended to be definitive, but are meant to act as a guide to inform decision-making. Information on other, perhaps rarer, highly predictive signs has been provided to help guide this decision. If the decision on evident toxicity is unclear then there is the option to continue the study to observe for the development of further clinical signs and/or test at a higher concentration if deemed necessary. However, we hope that this would be a rare occurrence, particularly as confidence grows in using this method.

References

Price, C., N. Stallard, S. Creton, I. Indans, R. Guest, D. Griffiths, and P. Edwards. 2011. A statistical evaluation of the effects of gender differences in assessment of acute inhalation toxicity. *Human & experimental toxicology*. 30:217-238.

Sewell, F., I. Ragan, T. Marczylo, B. Anderson, A. Braun, W. Casey, N. Dennison, D. Griffiths, R. Guest, T. Holmes, T. van Huygevoort, I. Indans, T. Kenny, H. Kojima, K. Lee, P. Prieto, P. Smith, J. Smedley, W.S. Stokes, G. Wnorowski, and G. Horgan. 2015. A global initiative to refine acute inhalation studies through the use of 'evident toxicity' as an endpoint: Towards adoption of the fixed concentration procedure. *Regulatory toxicology and pharmacology*. 73:770-779.

Stallard, N., C. Price, S. Creton, I. Indans, R. Guest, D. Griffiths, and P. Edwards. 2011. A new sighting study for the fixed concentration procedure to allow for gender differences. *Human & experimental toxicology*. 30:239-249.

Figures

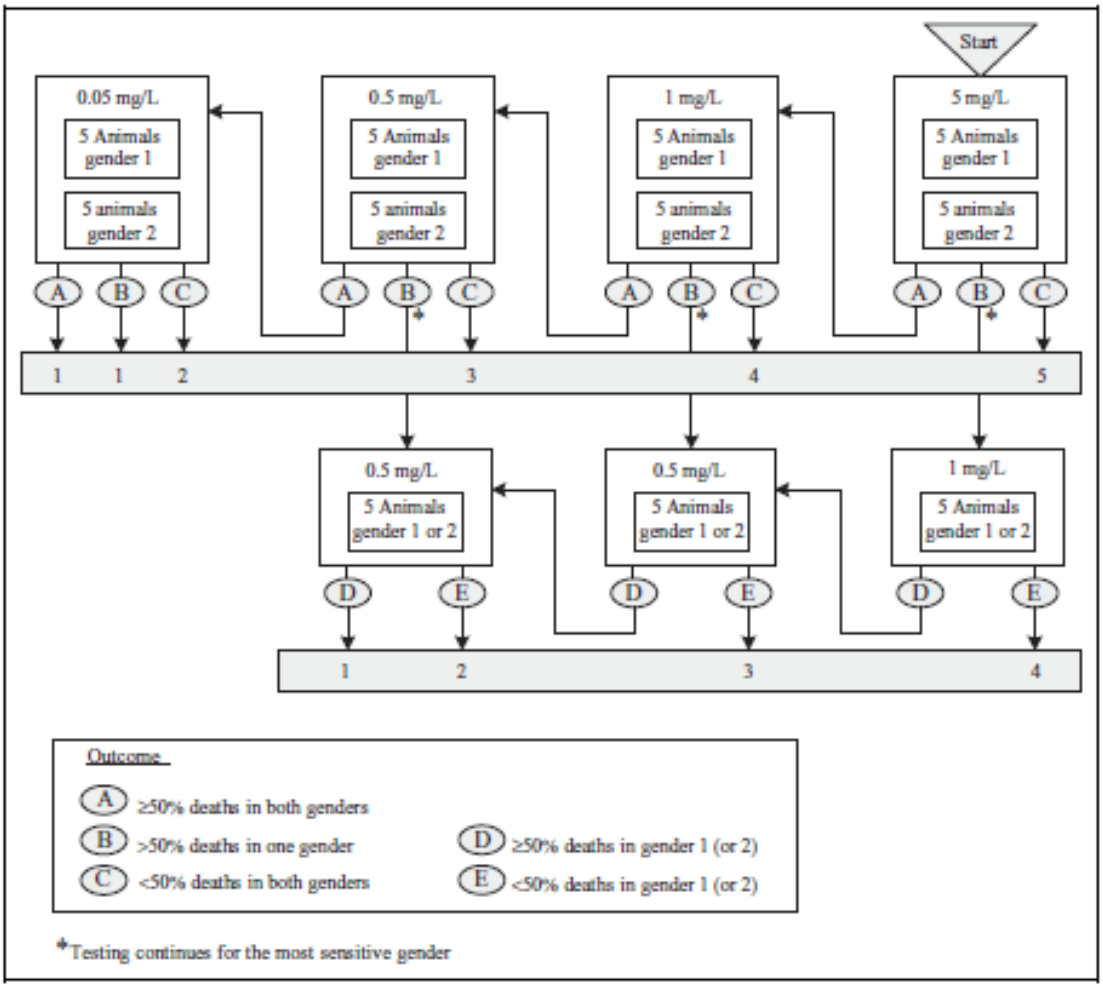


Figure 1: LC_{50} study protocol for dusts and mists starting at 5mg/L (Price *et al.*, 2011).

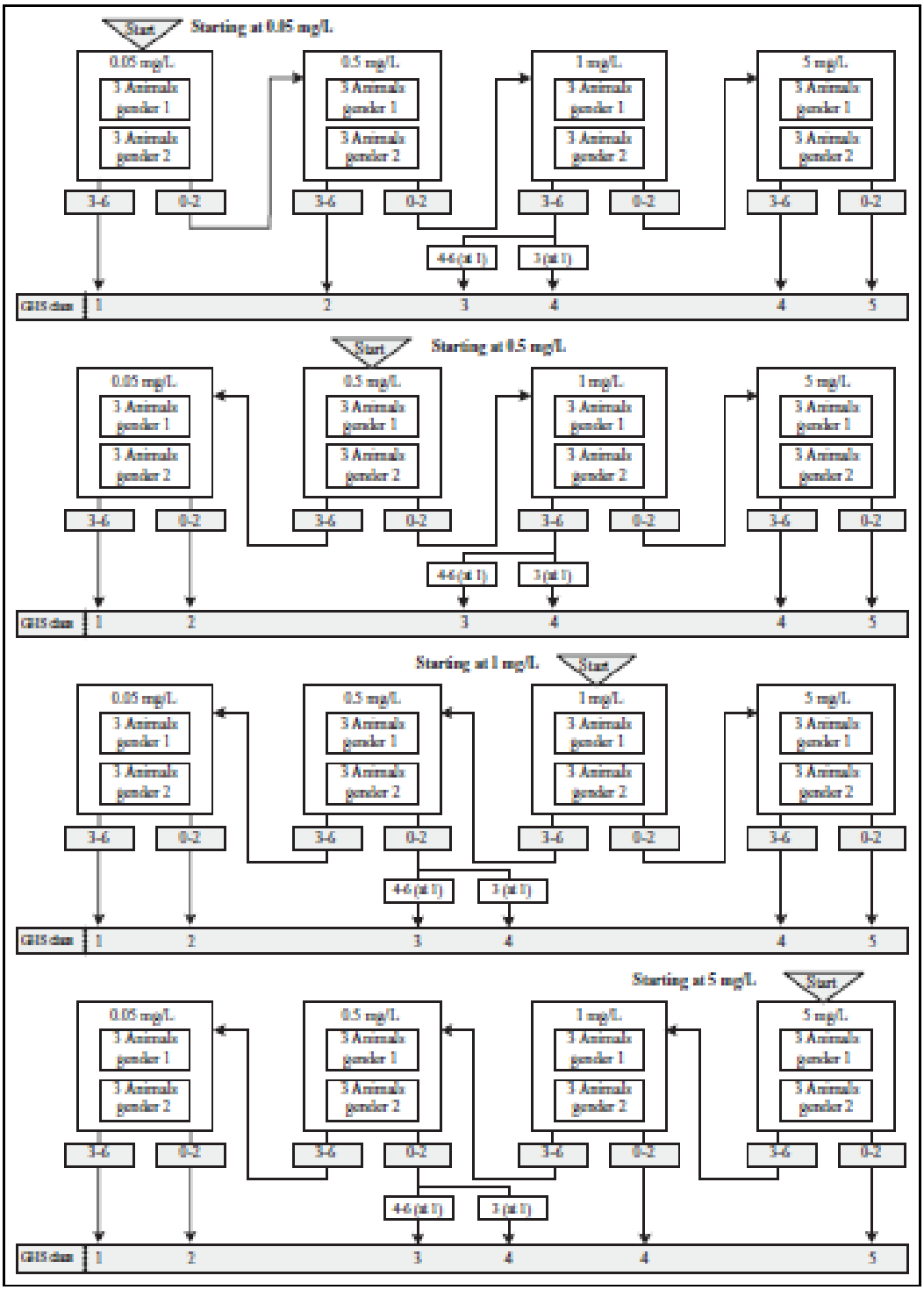


Figure 2: ATC study protocol for dusts and mists (Price *et al.*, 2011).

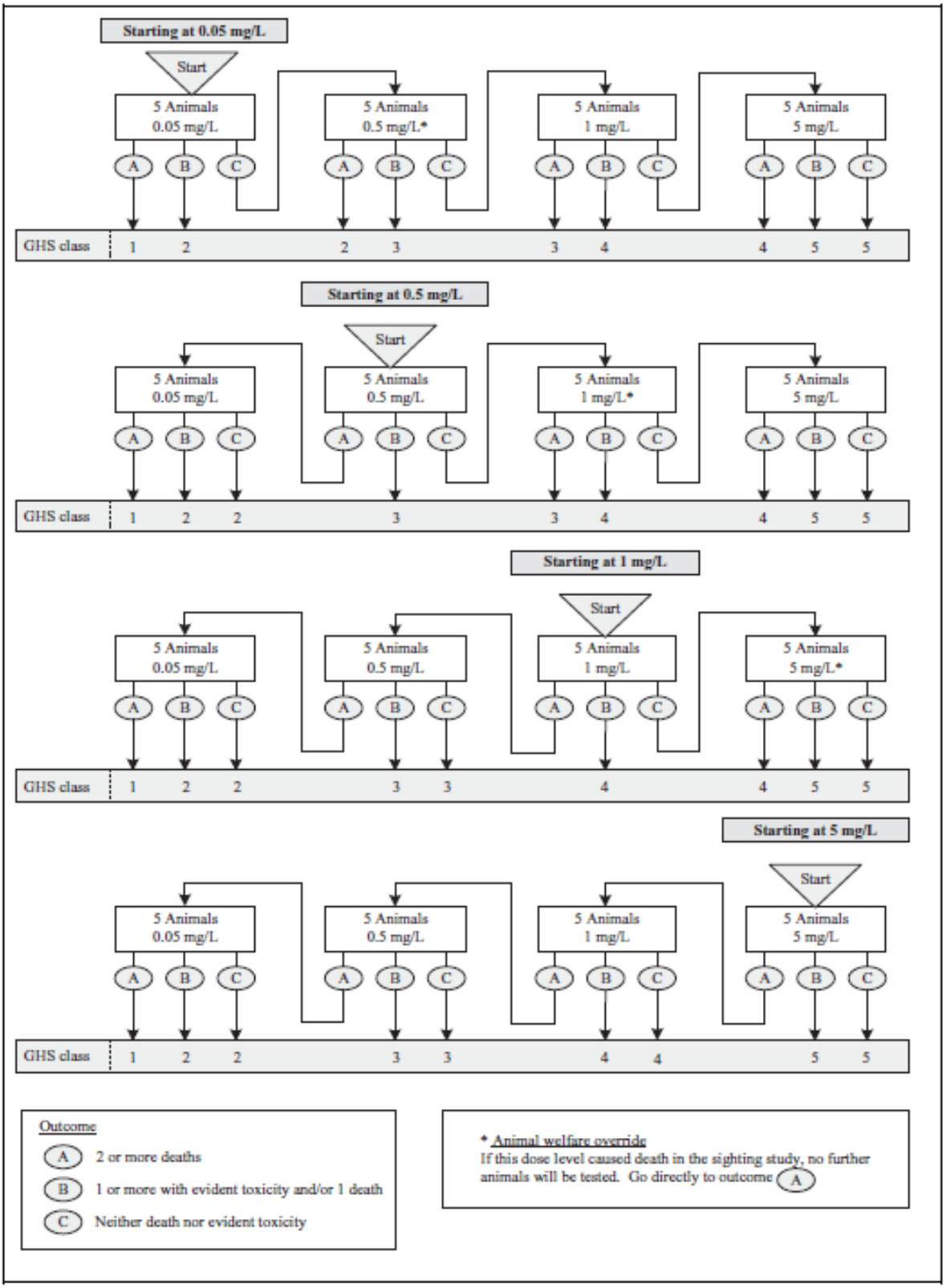


Figure 3: FCP main study protocol for dusts and mists (Price *et al.*, 2011).