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

01-Mar-2011

English - Or. English

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

OECD MRL CALCULATOR: STATISTICAL WHITE PAPER

Series on Pesticides
No. 57

JT03297201

ENV/JM/MONO(2011)3
Unclassified

English - Or. English

OECD Environment, Health and Safety Publications
Series on Pesticides

No. 57

OECD MRL Calculator
Statistical White Paper

INTER-ORGANIZATION PROGRAMME FOR THE SOUND MANAGEMENT OF CHEMICALS

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

Environment Directorate

ORGANISATION FOR ECONOMIC COOPERATION AND DEVELOPMENT

Paris 2011

Also published in the Series on Pesticides

No. 1 *Data Requirements for Pesticide Registration in OECD Member Countries: Survey Results* (1993)

No. 2 *Final Report on the OECD Pilot Project to Compare Pesticide Data Reviews* (1995)

No. 3 *Data Requirements for Biological Pesticides* (1996)

No. 4 *Activities to Reduce Pesticide Risks in OECD and Selected FAO Countries. Part I: Summary Report* (1996)

No. 5 *Activities to Reduce Pesticide Risks in OECD and Selected FAO Countries. Part II: Survey Responses* (1996)

No. 6 *OECD Governments' Approaches to the Protection of Proprietary Rights and Confidential Business Information in Pesticide Registration* (1998)

No. 7 *OECD Survey on the Collection and Use of Agricultural Pesticide Sales Data: Survey Results* (1999) [see also No.47]

No. 8 *Report of the OECD/FAO Workshop on Integrated Pest Management and Pesticide Risk Reduction* (1999)

No. 9 *Report of the Survey of OECD Member Countries' Approaches to the Regulation of Biocides* (1999)

No. 10 *Guidance Notes for Analysis and Evaluation of Repeat-Dose Toxicity Studies* (2000)

No. 11 *Survey of Best Practices in the Regulation of Pesticides in Twelve OECD Countries* (2001*)*

No. 12 *Guidance for Registration Requirements for Pheromones and Other Semiochemicals Used for Arthropod Pest Control* (2001)

No. 13 *Report of the OECD Workshop on Sharing the Work of Agricultural Pesticide Reviews* (2002)

No. 14 *Guidance Notes for Analysis and Evaluation of Chronic Toxicity and Carcinogenicity Studies* (2002).

No. 15 *Persistent, Bioaccumulative and Toxic Pesticides in OECD Member Countries,* (2002)

No. 16 *OECD Guidance for Industry Data Submissions for Pheromones and Other Semiochemicals and their Active Substances* (Dossier Guidance for Pheromones and other Semiochemicals) (2003)

No. 17 *OECD Guidance for Country Data Review Reports for Pheromones and Other Semiochemicals and their Active Substances* (Monograph Guidance for Pheromones and other Semiochemicals) (2003)

No. 18 *Guidance for Registration Requirements for Microbial Pesticides* (2003)

No. 19 *Registration and Work sharing, Report of the OECD/FAO Zoning Project* (2003)

No. 20 *OECD Workshop on Electronic Tools for data submission, evaluation and exchange for the Regulation of new and existing industrial chemicals, agricultural pesticides and biocides* (2003)

No. 21 *Guidance for Regulation of Invertebrates as Biological Control Agents (IBCAs)* (2004)

No. 22 *OECD Guidance for Country Data Review Reports on Microbial Pest Control Products and their Microbial Pest Control Agents* (Monograph Guidance for Microbials) (2004)

No. 23 *OECD Guidance for Industry Data Submissions for Microbial Pest Control Product and their Microbial Pest Control Agents* (Dossier Guidance for Microbials) (2004)

No. 24 *Report of the OECD Pesticide Risk Reduction Steering Group Seminar on Compliance* (2004)

No. 25 *The Assessment of Persistency and Bioaccumulation in the Pesticide Registration Frameworks within the OECD Region* (2005)

No. 26 *Report of the OECD Pesticide Risk Reduction Group Seminar on Minor Uses and Pesticide Risk Reduction* (2005)

No. 27 *Summary Report of the OECD Project on Pesticide Terrestrial Risk Indicators (TERI)* (2005)

No. 28 *Report of the OECD Pesticide Risk Reduction Steering Group Seminar on Pesticide Risk Reduction through Good Container Management* (2005)

No. 29 *Report of the OECD Pesticide Risk Reduction Steering Group Seminar on Risk Reduction through Good Pesticide Labelling* (2006)

No. 30 *Report of the OECD Pesticide Risk Reduction Steering Group: The Second Risk Reduction Survey* (2006)

No. 31 *Guidance Document on the Definition of Residue* [also published in the series on Testing and Assessment, No. 63] (2006, revised 2009)

No. 32 *Guidance Document on Overview of Residue Chemistry Studies* [also published in the series on Testing and Assessment, No. 64] (2006, revised 2009)

No. 33 *Overview of Country and Regional Review Procedures for Agricultural Pesticides and Relevant Documents* (2006)

No. 34 *Frequently Asked Questions about Work Sharing on Pesticide Registration Reviews* (2007)

No. 35 *Report of the OECD Pesticide Risk Reduction Steering Group Seminar on "Pesticide Risk Reduction through Better Application Technology"* (2007)

No. 36 *Analysis and Assessment of Current Protocols to Develop Harmonised Test Methods and Relevant Performance Standards for the Efficacy Testing of Treated Articles/Treated Materials* (2007)

No. 37 *Report on the OECD Pesticide Risk Reduction Steering Group Workshop "Pesticide User Compliance'* (2007)

No. 38 *Survey of the Pesticide Risk Reduction Steering Group on Minor Uses of Pesticides* (2007)

No. 39 *Guidance Document on Pesticide Residue Analytical Methods* [also published in the series on Testing and Assessment, No. 72] (2007)

No. 40 *Report of the Joint OECD Pesticide Risk Reduction Steering Group EC-HAIR Seminar on Harmonised Environmental Indicators for Pesticide Risk* (2007)

No. 41 *The Business Case for the Joint Evaluation of Dossiers (Data Submissions) using Work-sharing Arrangements* (2008)

No. 42 *Report of the OECD Pesticide Risk Reduction Steering Group Seminar on Risk Reduction through Better Worker Safety and Training* (2008)

No. 43 *Working Document on the Evaluation of Microbials for Pest Control* (2008)

Guidance Document on Magnitude of Pesticide Residues in Processed Commodities - only published in the Series on Testing and Assessment, No. 96 (2008)

No. 44 *Report of Workshop on the Regulation of BioPesticides: Registration and Communication Issues (2009)*

No. 45 *Report of the Seminar on Pesticide Risk Reduction through Education / Training the Trainers* (2009)

No. 46 *Report of the Seminar on Pesticide Risk Reduction through Spray Drift Reduction Strategies as part of National Risk Management* (2009)

No. 47 *OECD Survey on Countries' Approaches to the Collection and Use of Agricultural Pesticide Sales and Usage Data: Survey Results* (2009)

No. 48 *OECD Strategic Approach in Pesticide Risk Reduction* (2009)

No. 49 *OECD Guidance Document on Defining Minor Uses of Pesticides* (2009)

No. 50 *Report of the OECD Seminar on Pesticide Risk Reduction through Better National Risk Management Strategies for Aerial Application* (2010)

No. 51 *OECD Survey on Pesticide Maximum Residue Limit (MRL) Policies: Survey Results* (2010)

No. 52 *OECD Survey of Pollinator Testing, Research, Mitigation and Information Management: Survey Results* (2010)

No.53 *Report of the 1st OECD BioPesticides Steering Group Seminar on Identity and Characterisation of Micro-organisms* (2010)

No. 54 *OECD Survey on Education, Training and Certification of Agricultural Pesticide Users, Trainers and Advisors, and Other Pesticide Communicators: Survey Results* (2010)

No. 55 *OECD Survey on How Pesticide Ingredients Other than the Stated Pesticide Active Ingredient(s) are Reviewed and Regulated: Survey Results* (2010)

No.56 *OECD MRL Calculator User Guide, 2011*

Published separately

OECD Guidance for Country Data Review Reports on Plant Protection Products and their Active Substances-Monograph Guidance (1998, revised 2001, 2005, 2006)

OECD Guidance for Industry Data Submissions on Plant Protection Products and their Active Substances-Dossier Guidance (1998, revised 2001, 2005)

Report of the Pesticide Aquatic Risk Indicators Expert Group (2000)

Report of the OECD Workshop on the Economics of Pesticide Risk Reduction (2001)

Report of the OECD-FAO-UNEP Workshop on Obsolete Pesticides (2000)

Report of the OECD Pesticide Aquatic Risk Indicators Expert Group (2000)

Report of the 2nd OECD Workshop on Pesticide Risk Indicators (1999)

Guidelines for the Collection of Pesticide Usage Statistics Within Agriculture and Horticulture (1999)

Report of the [1st] OECD Workshop on Pesticide Risk Indicators (1997)

Report of the OECD/FAO Workshop on Pesticide Risk Reduction (1995)

About the OECD

The Organisation for Economic Co-operation and Development (OECD) is an intergovernmental organisation in which representatives of 34 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 ten different series: **Testing and Assessment**; **Good Laboratory Practice and Compliance Monitoring**; **Pesticides and 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;** and **Safety of Manufactured Nanomaterials.** 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/ehs/).

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, UNEP, UNIDO, UNITAR, WHO, World Bank and OECD. UNDP is an observer. 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

With the goal of harmonizing the calculation of MRLs across the OECD, the Residue Chemistry Expert Group of the OECD Working Group on Pesticides commissioned in 2008 an expert group to propose a new MRL calculation procedure. The guiding principles of this procedure were:

- the procedure must be a practical implementation of sound statistical methods;
- it must be simple to use without requiring extensive statistical knowledge from a user;
- it should produce a clear and unambiguous MRL proposal for most residue datasets produced by field trials; and,
- it should harmonize the EU and NAFTA procedures as much as possible.

The Working Group on Pesticides approved the draft OECD MRL Calculator and its User Guide in December 2010 and recommended that they be forwarded to the Joint Meeting of the Chemicals Committee and the Working Party on Chemicals, Pesticides and Biotechnology, for consideration as an OECD publication.

This document and the OECD MRL Calculator are being published under the responsibility of the Joint Meeting of the Chemicals Committee and the Working Party on Chemicals, Pesticides and Biotechnology, which has agreed that they be unclassified and made available to the public.

The OECD MRL Calculator is available on the OECD public website, http://www.oecd.org/env/pesticides under *Pesticide Publications*/*Publications on Pesticide Residues*.

TABLE OF CONTENTS

Introduction

1. There are two statistically-based calculation procedures in current use around the world for estimation of the MRL/tolerance from supervised field trial data sets: the so-called EU and NAFTA methods. The EU method has now been in use for a number of years in Europe and elsewhere. The NAFTA method, developed by a group of North American experts, has appeared recently and consequently has not been used as extensively as the EU method. However, both methods have come under criticism (see references below) and some commentators have highlighted apparent shortcomings in both methodologies.

2. With the goal of addressing those criticisms and harmonizing the calculation of MRLs across the OECD, the Residue Chemistry Expert Group (RCEG) during its meeting in Washington in 2008 commissioned an expert group formed by regulators and industry specialists to propose a new MRL calculation procedure. The guiding principles of this procedure are:

- the procedure must be a practical implementation of sound statistical methods;
- it must be simple to use without requiring extensive statistical knowledge from a user;
- it should produce a clear and unambiguous MRL proposal for most residue datasets produced by field trials; and,
- it should harmonize the EU and NAFTA procedures as much as possible.

3. Following these guiding principles, the OECD RCEG MRL calculation group began to work on the development and implementation of a robust methodology which was later considered to produce satisfactory results for the considerable number of real residue datasets tested.

4. The statistical goal of the OECD MRL Calculator, in common with previous methodologies, is to produce a MRL proposal in the region of the 95th percentile of the underlying residue distribution (which we abbreviate as p95), which is conservative in the sense that it will have a much greater propensity to make errors by overestimating p95 than by underestimating it for most datasets.

Residue Datasets

Field Trial Data Collection

5. Crop residue field trials (also referred to as supervised field trials) are conducted to determine the magnitude of the crop protection product residue in or on raw agricultural commodities, including feed items. In addition to studies for residues in crops grown in fields (i.e., outdoors), the OECD Crop Field Trials guidelines (see ref. [1]) also include studies to assess residues in protected crops grown in greenhouses (glass or plastic covering) and in crops treated after harvest (e.g., stored grains, wax or dip treatment of fruits). Residue field trials may have several objectives, such as: quantification of the expected range of residue(s) in crop commodities following treatment according to a particular good agricultural practice (GAP), determination of the rate of decline of residue(s) over time, determination of residue values such as the Supervised Trial Median Residue (STMR) and Highest Residue (HR) for conducting dietary risk assessment, or to provide data for the derivation of maximum residue limits (MRLs).

6. The critical GAP (*c*GAP) is generally used for residue field trials, this being the GAP leading to highest residues. Usually the *c*GAP would use the maximum number of applications at the maximum application rate and minimum re-treatment interval, with the shortest period between treatment and harvest of samples, whether defined by pre-harvest interval (PHI) or growth stage at application.

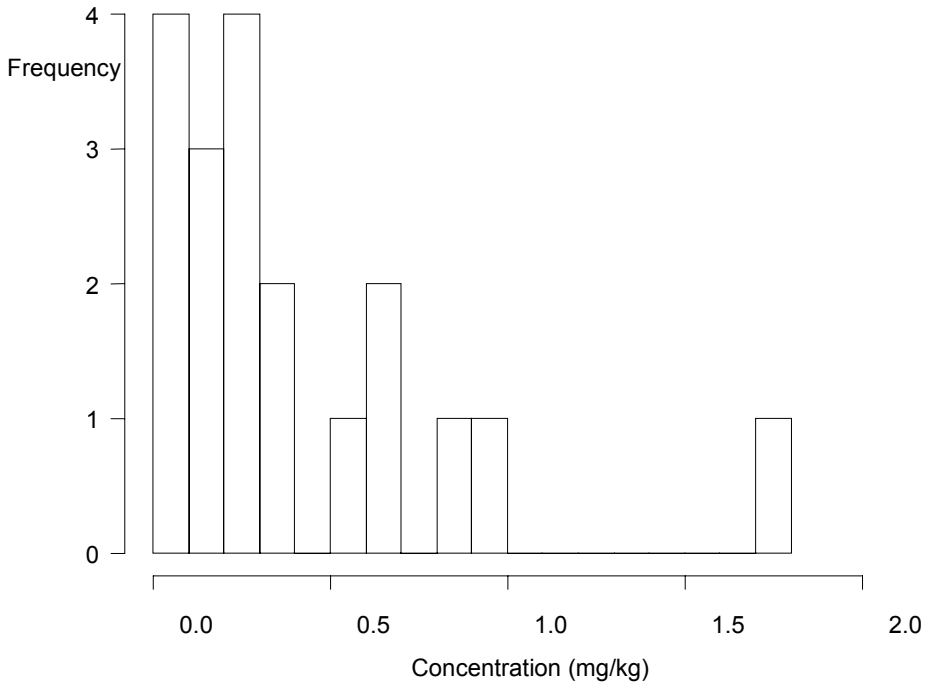
7. The current document is concerned with the calculation of MRLs, and it follows that the residue trials on which the calculation is to be based must have been analysed at least for those components included in the residue definition for enforcement. Processes for defining the residue are detailed in other guidance documents. It is also assumed for current purposes that the residue trials incorporate suitable quality-control procedures to ensure the reliability of the data produced.

8. Individual OECD countries or regions may have different requirements for the number, geographic distribution and type of residue field trials, or for the number of subsequent analyses. The design of a suitable residue trial program to satisfy these requirements can be complicated and is not discussed further here. The OECD MRL calculation procedure detailed in this document has been designed to accommodate most residue datasets arising from such a trial program.

Residue Data

9. The crop protection product (CPP) residue populations are usually left-censored (i.e. truncated at the limit of quantification (LOQ) levels), right skewed (i.e. asymmetric, having a long right tail) and contain extreme values that appear discrepant from the rest; see Figure 1.

Figure 1. Typical residue sample.



10. The censored values represent loss of information and can seriously affect the calculation of certain statistical measures like the mean and the standard deviation. The long right tail of the dataset leads to the appearance of residue values seven or eight times the size of the mean value. This complicates the classification of any of these extreme values as outliers.

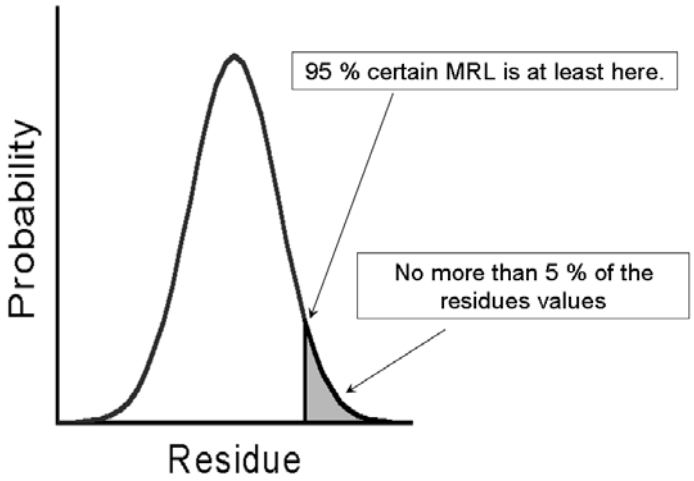
11. There is great diversity in the appearance of residue datasets. Some seem to follow a normal distribution (also called a bell-shaped or Gaussian distribution, see Figure 2). Others seem to follow a lognormal distribution (the logarithm of the residue values would follow the normal distribution), which is a right skewed distribution as depicted in Figure 1, especially when the coefficient of variation is large. Others still seem even more right skewed than the lognormal distribution and finally some residue datasets are so erratic that they do not appear to follow any known distribution at all. For a general introduction to environmental statistics see [2].

Previous Calculation Methods

EU Calculation

12. The EU gives detailed statistical guidelines for calculating MRLs (see [3]) which specify the use of two methods. Method I proceeds as if the samples were derived from a normal distribution and sets the MRL at the 95 % upper confidence limit (UCL) of the 95th percentile. That is, the MRL is set so that 95 % of the time, the 95th percentile of the assumed underlying normal distribution is lower than the MRL (see Figure 2). Both the confidence interval and the percentile are computed using the formulas corresponding to the normal distribution.

13. Figure 2. In EU method I a normal distribution is assumed.



14. Method II does not assume that the residue data follow any particular distribution. Instead, the 75th percentile of the sample is computed and then doubled. The percentile is computed using the Weibull procedure (see [4]).

15. It is important to point out that the "PERCENTILE" function available in Excel is not appropriate for computing the percentiles mentioned in Method I or Method II. For Method I, the appropriate methodology derived from the normal distribution should be used as explained in [3]. For Method II, a special Excel add-in should be produced, which implements the Weibull method as described in [4].

16. The EU regulations do not provide guidance for when to use Method I or Method II or for what to do when the MRLs produced by the two methods differ. It simply says that the next step consists in the rounding of the MRL value to one of 16 discrete MRL classes listed in [3].

17. The EU guidelines require the substitution of non-detected residues by the LOQ values. This give a much exaggerated worst-case scenario and it is clearly undesirable from a statistical point of view because it skews the residue distributions, inflates the estimator of the mean and decreases the estimators of the variability [5, 6]. On the other hand, the guidelines allow for the removal of suspected outliers by using the Dixon's Q-test, which assumes normality of the residue population. The guidelines warn very appropriately against the use of Dixon's Q-test for non-normal residue distributions.

18. Of the two methods included in the guidelines, Method I is sensitive to the substitution of non-detects and the removal of outliers, due to the fact that it is based on a normality assumption. Method II is not sensitive (for a small proportion of non-detects), because it does not make any distributional assumption. See reference [7] for a performance evaluation of this methodology.

NAFTA Proposed Calculation

19. Recently, a new method for MRL calculations has been proposed for the NAFTA area, where MRLs may also be referred to as "tolerances" [8]. During the development process, the regulators considered a number of possible calculation methods and selected some of them for use.

20. The first part of the procedure requires "filling in" the values of the non-detects (NDs, which stands here for LOQs) by assuming that the samples have been produced from a lognormal distribution. The log-normality of the resulting dataset is checked both by the use of the Shapiro-Francia test as well as by a visual inspection. If the dataset is considered not to be lognormal, then the MLR value is set three standard deviations above the sample mean (for a normal distribution, this would be roughly equivalent to setting the MRL above the 99th percentile).

If the dataset is deemed lognormal, then up to three different statistical measurements may be required:

- the 95% upper confidence limit on the 95th percentile;
- the 99th percentile estimate and
- the product of 3.9 and the upper prediction limit of the median (this quantity is referred to as "UCLMedian95").

21. All these measures are calculated following the rules of the lognormal distribution; so although the first measure looks very similar to the one used in the EU Method I, it is likely to produce a higher result. The third measure is produced under the additional assumption that the *coefficient of variation* CV (the ratio of the standard deviation to the mean) has a value of one.

22. For large datasets (more than 15 data points), the minimum of the first two measures is taken forward (these two measurements are referred collectively as the "95/99 rule"). For smaller datasets, the minimum of all three measures is required. Whichever option is chosen, the result is rounded up according to a set procedure within the calculator.

23. There is no allowance in the NAFTA procedure for the removal of suspected outliers by any statistical method. See reference [9] for a performance evaluation of this methodology.

Evolution of the OECD MRL Calculator

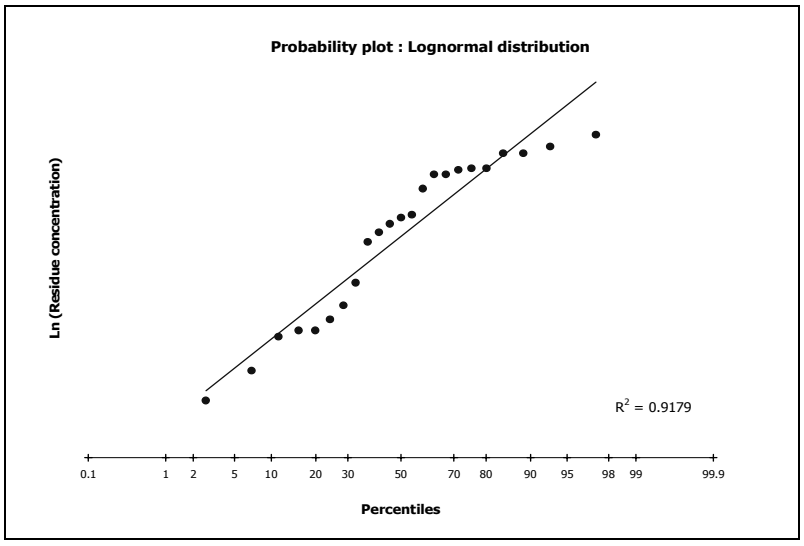
24. The initial draft versions of the OECD Calculator employed a distributional analysis similar to the methodology used in the NAFTA Calculator. In addition to consideration of a lognormal distribution, the data were also evaluated against a normal and Weibull distribution. The distribution with the highest correlation coefficient was selected. This distribution was used to calculate a MRL proposal as the minimum of a) the 95% upper confidence limit of the 95th percentile and b) the point estimate of the 99th percentile. For residue data sets that were too small for distributional analysis or failed to fit any of the three distributions a MRL proposal was calculated from the non-distributional approach, i.e. the Mean + 3*SD method.

25. In these earlier versions of the OECD Calculator, the regulatory ceiling was introduced as a means of preventing MRL proposals greater than 2 X highest residue or 3 X median. Experience with the NAFTA Calculator and lognormal distributional methods, showed that such relatively high MRL proposals in relation to the highest residues often occurred when residue data did not fit the distribution well at the

upper end ("tailing effect"). It was initially thought by the group that the ceiling was a pragmatic approach which could be justified by common sense and historical experience, not statistics. It was always recognized that the ceiling was arbitrary and not supported by any statistical justification. For this reason, the regulatory ceiling was identified as an aspect of the Calculator which required focused testing and evaluation.

26. Early testing indicated that the regulatory ceiling itself was selected as the MRL in only one of 482 EU data sets tested; however, the regulatory ceiling influenced the selection of the distribution in 2% of the data sets tested. Focusing on this effect, the individual data sets that were influenced by the regulatory ceiling were reviewed to see if the regulatory ceiling was in some statistical or pragmatic way "sensible" or completely arbitrary and unjustified. In all cases these individual data sets showed a tailing effect in the probability plot which resulted in a larger than expected MRL when compared to the highest residue. An example is given below.

Figure 3. Example of suspected "tailing".



27. This tailing effect was observed in a small, but noticeable percentage of the data sets analyzed using the NAFTA Calculator, usually resulting in relatively high MRLs. With the three different distributions employed in the older version of the OECD Calculator, this tailing effect was seen less often. Nevertheless, since a small percentage of data still showed this tailing effect, an effort was made to find or develop a statistical fit test for tailing which could be used in the Calculator to replace the regulatory ceiling. Unfortunately, no test could be found which worked as reliably with residue data as the regulatory ceiling.

28. In today's version of the OECD Calculator, distributional tests are no longer employed (see appendix B for details). As a result, "tailing" is no longer an issue which needs to be addressed. In addition, the incidence of unusually high MRLs relative to the residue population is decreased using the current non-distributional approach. In Figure 9 below which shows the results of testing real data sets, the MRLs proposed from the current version of the OECD Calculator are generally below 2 x HR for data set sizes of 20 and 16 data points. Testing using synthetic data was even more convincing, showing that 95% of the MRL proposals using the current OECD Calculator were at or below 2 x HR for data set sizes of 10 or more data points (see Figure 5). The current methodology of the OECD Calculator does not restrict the MRL proposal in relation to either the highest residue or the median of the data set. In fact, for smaller data

sets (i.e., less than 10 points), there is a possibility that the MRL proposal will exceed 2, or even rarely, 3 times the HR. For these dataset sizes, where the uncertainty is inherently high, this is considered justified, especially when there is a great deal of variability within the data set.

The Current OECD MRL Calculator

29. The current OECD Calculator distinguishes fully censored residue datasets (sample sets with all measurements below one or several limits of quantification) from not fully censored datasets (datasets with at least one measurement at or above the LOQ of the corresponding analytical method).

For field trials where sampling replicates were taken (more than one composite sample for that field trial), the calculator group recommends using the average or mean value of the replicates as the representative value for that field trial in exactly the same fashion that is done for analytical replicates of the same composite sample. From a statistical point of view, the mean or average residue value of replicate samples provides the basis for setting MRLs targeted at the p95 of the underlying distribution (see Performance of the OECD Calculator below). However, there may be situations where single valid results from replicate samples may exceed the MRL estimated from the use of average or mean values. In such situations and in view of consumer safety, consideration may be given by some regulatory authorities to the use of these single values as the HR in dietary risk assessment.

30. Please see appendix D for a justification of this practice and a discussion of a different recommendation made by the JMPR committee in their 2007 Report.

Not Fully Censored Datasets

31. For not fully censored datasets, the maximum of three calculated results is put forward as the MRL proposal by the calculator:

- the highest residue is used as a "floor" to guarantee that the MRL proposal is always greater than or equal to the highest residue[1];
- the mean and the standard deviation values of the dataset are computed; the "mean + 4* standard deviation" value is evaluated as the base proposal (referred to as "Mean + 4*SD" method); and,
- the "3*Mean*CF" method (see next paragraph).

Note: for the calculation of the mean and standard deviation, all values less than the LOQ are to be introduced into the calculator with a value equal to the LOQ.

32. The "3*mean" value is computed to provide another "floor" to the calculation; in this case to guarantee that the sample coefficient of variance (CV = standard deviation / mean) used in the calculation is at least 0.5, a condition verified by most residue datasets[2]. This is necessary given the tendency of small datasets to underestimate the standard deviation[3]. A correction factor CF has been added because it was

[1] This requirement was introduced from feedback obtained through the circulation for over a year in various JMPR and OECD committees of a questionnaire that contained policy questions needed to complete the design of the MRL calculator.

[2] If the CV = 0.5, then "SD = 0.5*Mean" and then we have that "Mean + 4*SD" = "Mean + 2*Mean" = "3*Mean."

[3] In a previous version of the calculator, the "3*Mean*CF" method was only applied to datasets of 15 or less data points. But after considerable search, the calculator group was unable to find an example of a datasets of

observed that the mean of a dataset is overestimated for censored datasets. The correction factor CF is equal to 1 – ⅔*fraction censored data in the dataset. This calculation is referred to as the "3*Mean*CF" method.

So the MRL proposal for not fully censored datasets is,

Maximum (Highest Residue, Mean + 4*SD, 3*Mean*CF).

33. The case of almost fully censored datasets but with several LOQ value is more complicated, especially when there are quantified values below the largest LOQ value. The above procedure is still used and will produce an MRL proposal but the user may consider reviewing this proposal on a case-by-case basis.

Fully Censored Datasets

34. The *OECD Residue Chemistry Expert Group* (RCEG) proposed in their August 2010 meeting in Washington that the MRL for fully censored datasets be set at the level of the highest LOQ present in the dataset. In doing so, that committee decided not to adopt a recommendation made by the calculator group. Details of the calculator group original proposal are given in Appendix C.

35. Consideration was given to the proposal of using residue values that are below the LOQ of a method, but above the LOD (limit of detection), especially as a refinement in cases where detectable residue values below the LOQ are available. It was decided that this distinction would not be included in this calculator due to the inherent challenges associated with these residue values (they are not supported by validation data, they may be very close to the limit of detection of the instrumentation and they are often not available to regulators). Nevertheless, considerable testing was conducted with these <LOQ residue values to see the effect on proposed MRLs. Fortunately, due to the robustness of the "Mean + 4*SD" method and the inclusion of the censoring factor (CF) for the "3*Mean" method, censoring does not strongly influence the MRL proposal; therefore, including residue values below the LOQ usually affects the MRL proposal little, if at all.

Rounding

36. To facilitate the setting of harmonized MRLs in the global environment, MRL proposals are rounded as a last step in the calculation. For numbers between 1 and 10, they are rounded to a single digit; for 10 to 100, they are rounded to multiples of 10; for 100 to 1000, they are rounded to multiples of 100 and so on. Intermediate values of 0.015, 0.15, 1.5, 15, etc, were introduced to avoid doubling of MRLs on rounding. So for example: 0.12 rounds up to 0.15, 0.16 rounds up to 0.2; and 12 rounds up to 15 instead of 20. The possibility for rounding down exists if a particular MRL level is surpassed by a specified amount.

more than 15 non-censored data points with a CV of less than 0.5, so the distinction between small and larger datasets was removed from the calculator to favor simplicity. If an applicant is in possession of a large dataset of non-censored residue values with a CV smaller than 0.5, the applicant may put forward a case for dropping this "3*Mean*CF" method from the calculation of that particular MRL.

To be more precise, the rounding possibilities are (in mg/kg):

0.001	0.0015	0.002	0.003	0.004	0.005	0.006	0.007	0.008	0.009
0.01	0.015	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.1	0.15	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
1	1.5	2	3	4	5	6	7	8	9
10	15	20	30	40	50	60	70	80	90
100	150	200	300	400	500	600	700	800	900
1000	...								

37. If it is not desired to set MRLs below 0.01 mg/kg, smaller MRL proposals may be rounded up to that value. If the 0.015 mg/kg is not desirable due to limitations in the analytical methods, the MRL may be rounded up to 0.02 mg/kg.

38. MRLs are displayed without decimal zeroes after the last significant figure, to avoid giving the impression of having more accuracy than in reality. So, for example, a MRL is displayed as 2 mg/kg but not 2.0 mg/kg; 0.1 mg/kg is possible but 0.10 mg/kg is not.

39. Rounding down will happen if the MRL proposal exceeds the lower MRL rounding possibility by less than 10% of the difference between the upper and lower MRL rounding possibilities. For example:

MRL Class	10% of Difference	Cut off Point for Rounding Down
0.02	0.001	0.021
0.03	0.001	0.031
...	...	....
0.09	0.001	0.091
0.1	0.005	0.105
0.15	0.005	0.155
0.2	0.01	0.21
0.3	0.01	0.31
...	...	...
0.9	0.01	0.91
1	0.05	1.05
1.5	0.05	1.55
2	0.1	2.1
3	0.1	3.1

Some rounding examples:

Unrounded proposal:	1.04 mg/kg	→	Rounded proposal: MRL:	1 mg/kg
Unrounded proposal:	1.12 mg/kg	→	Rounded proposal: MRL:	1.5 mg/kg
Unrounded proposal:	1.53 mg/kg	→	Rounded proposal: MRL:	1.5 mg/kg
Unrounded proposal:	1.58 mg/kg	→	Rounded proposal: MRL:	2 mg/kg
Unrounded proposal:	2.07 mg/kg	→	Rounded proposal: MRL:	2 mg/kg
Unrounded proposal:	2.12 mg /kg	→	Rounded proposal: MRL:	3 mg/kg
Unrounded proposal:	21.0 mg/kg	→	Rounded proposal MRL:	30 mg/kg

Performance of the OECD Calculator

40. The procedure described above was chosen by the calculator group since a) it is simple to use, b) it does not depend on distributional assumptions[4], c) it is robust in relation to the presence of censored data, and d) it performs better compared to the distributional methods, especially for small datasets. The performance of the procedure was tested on both synthetic and real datasets. Also, MRL proposals were compared with historical MRLs of EFSA and JMPR, as well as with the MRLs produced by the NAFTA Calculator.

Performance against synthetic data

41. Testing on synthetic datasets was performed with 100,000 datasets sampled from the lognormal distribution with the CV = 1.0[5]. This distribution was believed to represent a reasonable worst case for real field trial data, which means that performance is expected to be better than depicted below for most datasets.

42. For each dataset generated from the lognormal distribution, a MRL proposal was calculated. For the smallest dataset that we consider (with at least 3 data points), most of the calculated MRL proposals (95%) lay between 0.37*p95 and 4.50*p95 (Figure 4), while the MRL-over-HR ratio varied between 2.0 and 2.7 (Figure 5). We call these intervals the 95% probable ranges of the computation, because 95% of the time the results are within that range. The failure rate, i.e. the chance to get a MRL below the p95, was about 42.5% for 3 data points (Figure 6). It decreased to approximately 25% for 8 data points, and the level of 5% was reached for 29 data points.

[4] Although this method does not depend on any distributional assumption, a classical theorem called the Chebishev's inequality states that for any large enough sample extracted from any distribution with finite mean and variance, the "Mean + 4*SD" method will provide an estimate for a percentile above the 93th percentile. For the lognormal distribution, it will provide estimates above the 99th percentile for large enough samples.

[5] For the underlying normal distribution, the mean was 1 and the standard deviation was $\ln(2)^{1/2} \sim 0.83$.

Figure 4. 95% probable ranges for the MRL-over-p95 ratio depending on the number of data points sampled from a lognormal distribution with the CV of 1. Red (+) and green (*) labels show the upper and lower boundaries, respectively. The blue line represents the equality of the ratio to one. The right axis displays the corresponding percentile of the lognormal distribution. Ratios are in a linear scale on the left graph and in a logarithmic scale on the right graph. This example represents a reasonable worst case for real field trial data, which means that performance is expected to be better than depicted above for most real datasets. Rounding will move both green and red lines up.

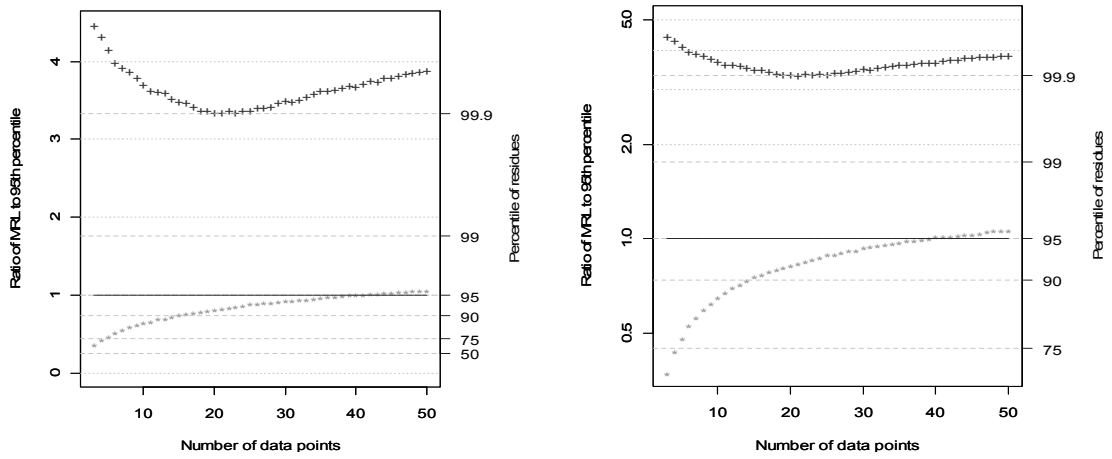


Figure 5. 95% probable ranges of the MRL-over-HR ratio depending on the number of data points. Red (+) and green (*) plus signs show the upper and lower boundaries, respectively. This example represents a reasonable worst case for real field trial data, which means that performance is expected to be better than depicted above for most real datasets. Rounding will move green and red lines up.

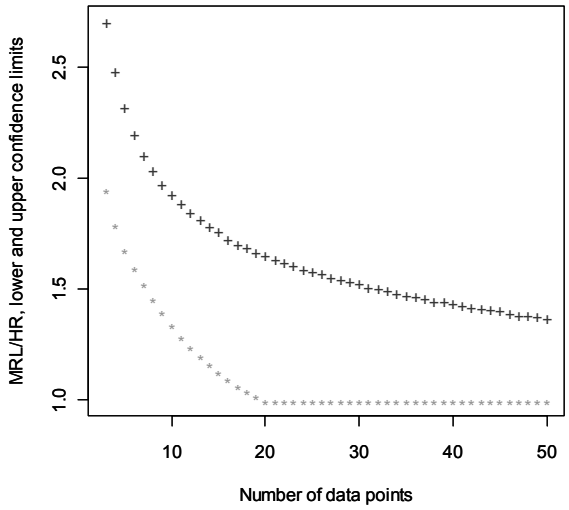
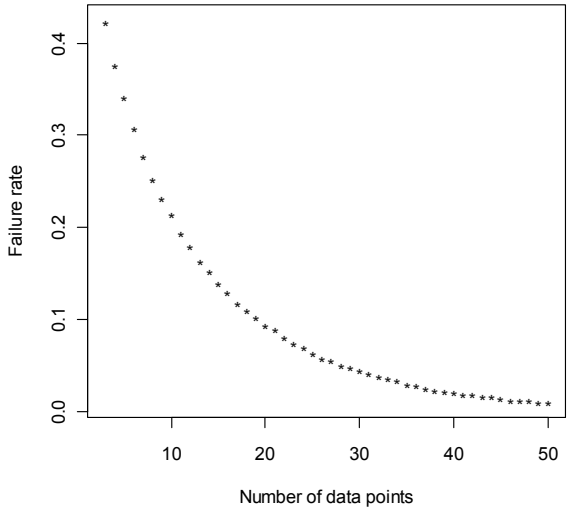


Figure 6. "Failure rate", i.e. a fraction of datasets for which the proposed MRL is below the p95, versus the number of data points. This example represents a reasonable worst case for real field trial data, which means that performance is expected to be better than depicted above for most real datasets. Rounding will reduce the failure rate.



43. So even for this reasonable worst case and for an extremely small dataset, the MRL proposal is expected to be roughly between half and four times the 95th percentile. Even for the extremely small datasets, there is a greater probability that the MRL proposal will be above the 95th percentile instead of below it (almost 60% confidence).

Performance against real data (sub-sampling from large datasets)

44. A considerable amount of effort and time was dedicated to compare the performance of the different calculation options considered during the course of this project against real datasets. In this paper we will only report the results for the suggested method.

45. For a real dataset, the underlying distribution as well as any of its percentiles is unknown. To evaluate the performance against real datasets of any methodology that was under consideration, we selected large real datasets (at least 20 or 30 residue values), extracted multiple subsets of different sizes from them (without replacement), and computed MRL proposals for all these subsets.[6] The HR of the large dataset is expected to be in the range of the highest percentiles of the underlying distribution (typically above the 95th percentile for a dataset of more than 20 residue values); so it can be taken as a reference point, as done in Figure 7.

46. As can be seen in that figure, for subsets of 16 and 20 points, most of the MRL proposals are above the HR of the full dataset. For subsets of 5 and 8 points, most MRL proposals lie between 0.5 and 2.5 of the HR of the parent dataset. Even for these small subsets, the MRL proposal is above the HR of the parent dataset much more frequently than below.

47. Another way of evaluating the performance (and also the robustness) of the calculator is displayed in Figure 8. The ratio of the MRL proposal for the subset over the full dataset is displayed. For all subset sizes, there is a very clear peak around 1, which means a perfect match. For very small subsets, the peak ratios level off, producing values between 0.5 and 1.5. Finally in Figure 9, the ratio of the MRL proposal of a subset to its HR is shown to illustrate that for very small datasets, it may even be above 3.

48. For these tests, 10 subsets of 5, 8, 16 and 20 points were extracted from each of 63 large datasets obtained from published EFSA residue data and JMPR data.

[6] We sample without replacement, so this is similar to bootstrapping but not identical.

Figure 7. Ratio of the MRL proposal for each subset and the HR of the full set after rounding.

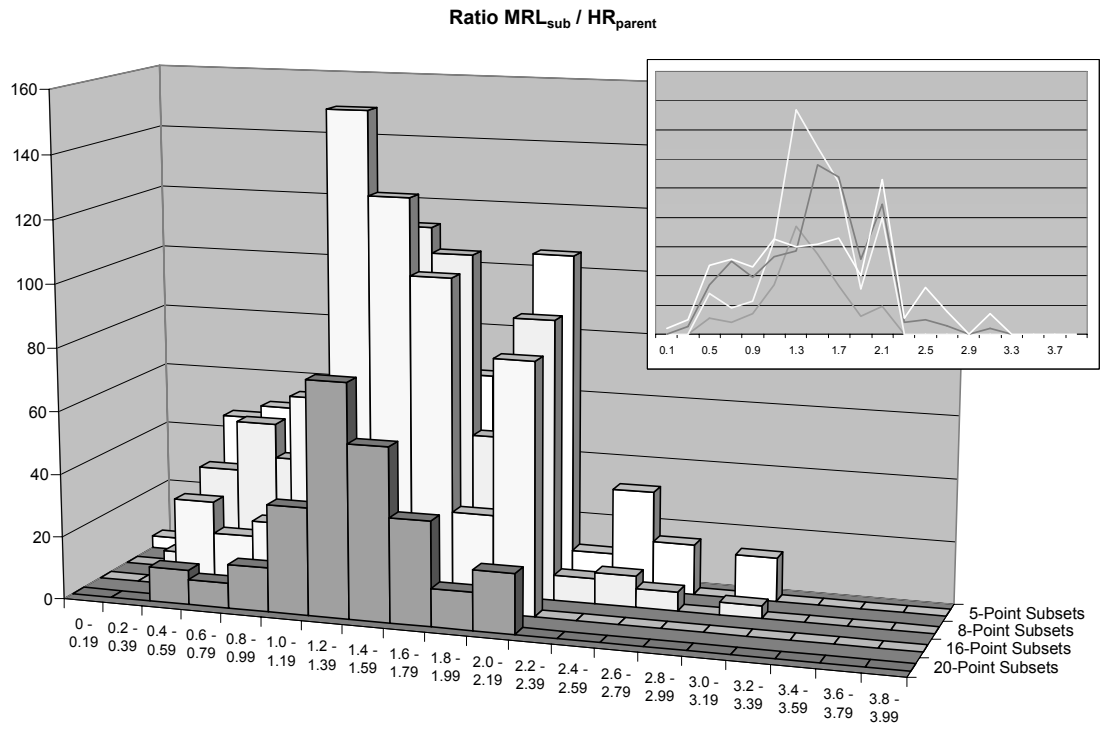


Figure 8. Ratio of the MRL proposal for the subset and the MRL proposal for the full set after rounding.

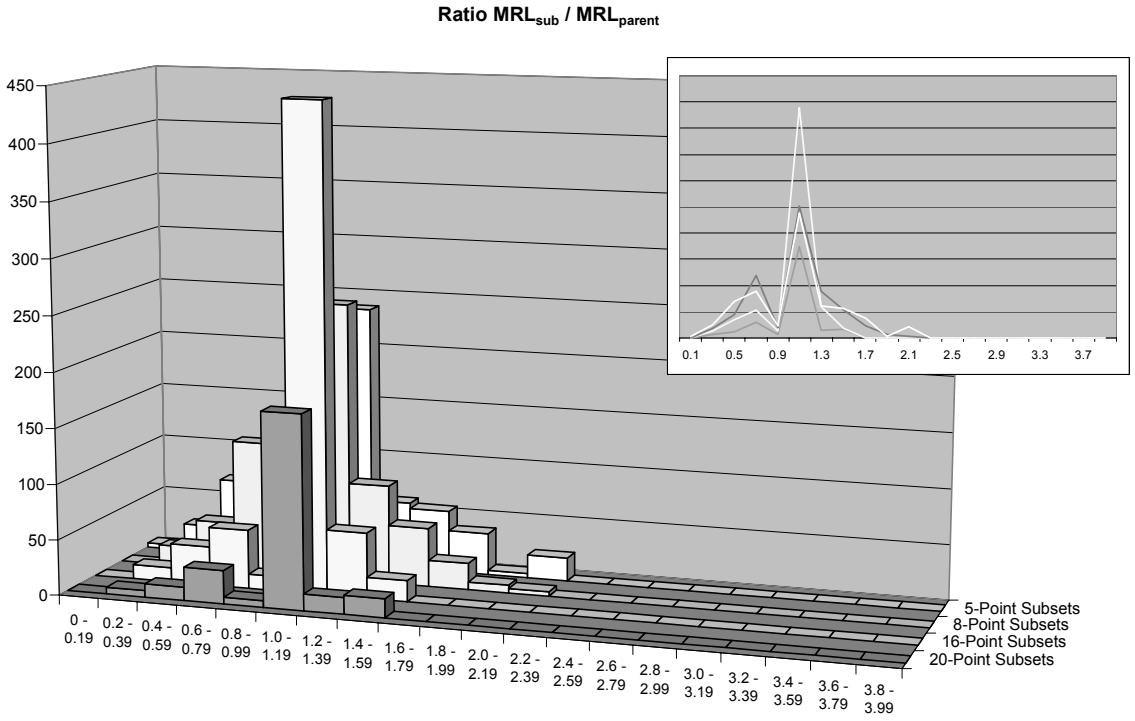
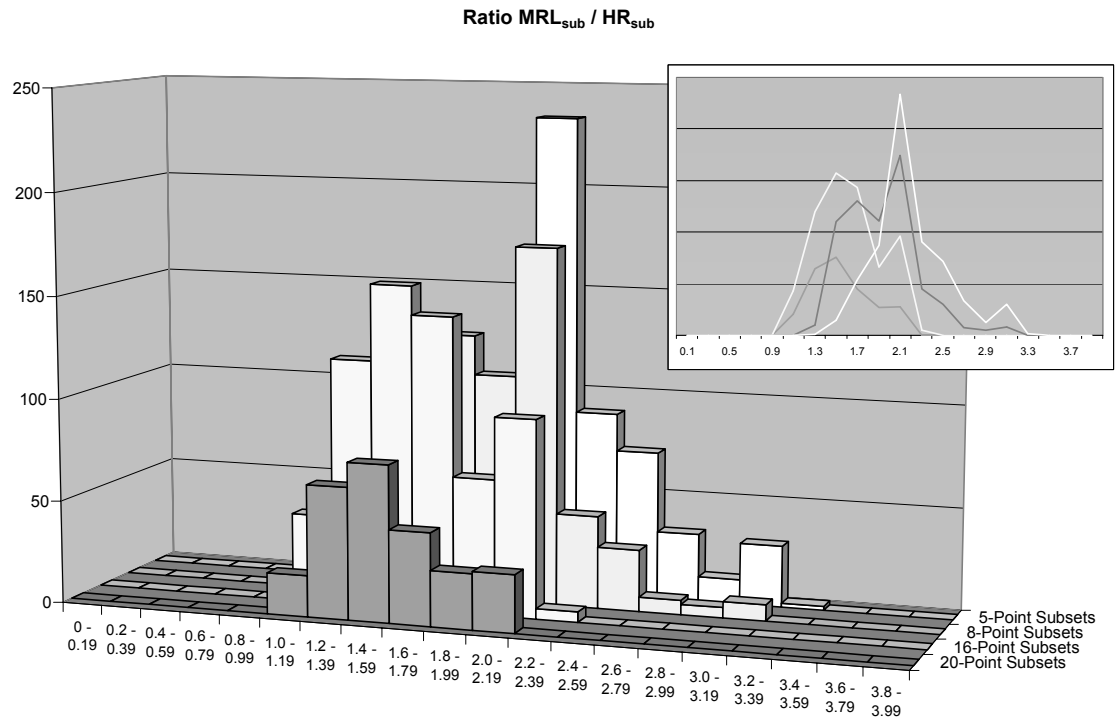


Figure 9. Ratio of the MRL proposal of a subset to its HR after rounding.



Comparison with historical MRLs

49. The proposed method was tested using real residue data that were recently evaluated by experts of EFSA or JMPR and used by these experts to derive MRLs. The residue data evaluated by EFSA were taken from "Reasoned Opinions on MRLs" published by EFSA between November 2008 and February 2010. The residue data evaluated by JMPR were taken from the JMPR Report 2008. Some of the characteristics of the two residue data collections are shown below.

Table 1. Characteristics of EFSA and JMPR residue data collections.

	EFSA	JMPR
Total number of datasets	215	201
Number of different active substances	47	15
Number of different commodities	ca. 70	ca. 80
Number of datasets with $n \leq 4$	52	17
Number of datasets with $4 < n \leq 8$	110	79
Number of datasets with $8 < n \leq 16$	44	62
Number of datasets with $n > 16$	9	43
Number of datasets without any censored value	148	120
Number of fully censored datasets	14	15

50. In case of real datasets, where the underlying distributions are not known, it is difficult to assess whether the MRLs proposed by the calculator are suitable or not. Thus the performance of the calculator was characterized by doing some statistics on:

- the ratio between the rounded MRL and the HR;
- the ratio between the rounded MRL and the unrounded MRL;
- the ratio between the rounded MRL and the MRL proposed by EFSA or JMPR; and,
- the ratio between the rounded MRL and the MRL derived from the same dataset using the NAFTA Calculator (note: when using the NAFTA Calculator censored data were not replaced by maximum likelihood estimates; that is, the MLE procedure was not used).

51. For each of these ratios the minimum, maximum and mean were determined for all the EFSA and all the JMPR datasets. Furthermore the mean was also calculated by grouping the datasets according to size and percentage of censored data (see Table 2 and Table 3).

Table 2. Test with the proposed MRL Calculator using EFSA residue data.

EFSA datasets	Rounded MRL / Unrounded MRL	Rounded MRL / HR	Rounded MRL / EFSA MRL	Rounded MRL / NAFTA calc. MRL
Min (overall)	0.95	1.00	0.30	0.43
Max (overall)	4.16	10.31	2.00	2.67
Mean (overall)	1.19	2.05	1.12	1.11
Mean ($n \leq 4$)	1.30	2.54	1.08	1.16
Mean ($4 < n \leq 8$)	1.18	2.02	1.12	1.09
Mean ($8 < n \leq 16$)	1.12	1.68	1.21	1.08
Mean ($n > 16$)	1.07	1.35	0.92	1.08
Mean (0% censored)	1.11	2.03	1.16	1.12
Mean (< 100% censored)	1.21	2.12	1.14	1.11
Mean (100% censored)	1.01	1.01	0.91	0.98

Table 3. Tests with the proposed MRL Calculator using JMPR residue data.

JMPR datasets	Rounded MRL / Unrounded MRL	Rounded MRL / HR	Rounded MRL / JMPR MRL	Rounded MRL / NAFTA calc. MRL
Min (overall)	0.96	1.00	0.40	0.50
Max (overall)	1.39	3.13	3.00	3.00
Mean (overall)	1.10	1.78	1.05	1.15
Mean (n ≤ 4)	1.09	2.35	1.28	1.37
Mean (4 < n ≤ 8)	1.08	1.96	1.08	1.12
Mean (8 < n ≤ 16)	1.10	1.62	1.00	1.11
Mean (n > 16)	1.13	1.47	0.95	1.20
Mean (0% censored)	1.09	1.95	1.05	1.16
Mean (< 100% censored)	1.11	1.85	1.05	1.17
Mean (100% censored)	1.00	1.00	1.04	1.00

52. On average the rounding procedure tends to increase the MRLs by about 10—20%. With some EFSA datasets, a more than 3-fold increase due to rounding was observed. this is because the LOQ for these datasets was 0.00097 mg/kg and the unrounded MRLs were far below the lowest MRL class of 0.01 mg/kg implemented in the OECD Calculator.

53. The mean ratio between the rounded MRL and the HR was 2.1 and 1.8 for the EFSA and JMPR dataset collections, respectively. This ratio tends to decrease when the size of the dataset increases. A more than 10-fold ratio was observed for some of the EFSA datasets that were analysed with an LOQ of 0.00097 mg/kg and which contained no data higher than the LOQ.

54. On average the MRL estimates yielded by the OECD Calculator exceed the MRLs proposed by EFSA and JMPR experts by 12% and 5%, respectively. However, larger deviations are observed for some individual datasets. The ratio between the MRL estimates produced by the OECD Calculator and the MRLs proposed by experts ranges between 0.30 and 2.0 for the EFSA datasets and between 0.40 and 3.0 for the JMPR datasets.

55. On average the MRL estimates yielded by the OECD Calculator also tend to exceed the MRLs yielded by the NAFTA Calculator. The average exceedance is 11% for the EFSA datasets and 15% for the JMPR datasets. Again, larger deviations are observed for some individual datasets. The ratio between the MRL estimates produced by the OECD Calculator and the MRLs produced by the NAFTA Calculator ranges between 0.43 and 2.7 for the EFSA datasets and between 0.50 and 3.0 for the JMPR datasets.

56. The following graphs allow for comparison of the MRLs produced by the OECD Calculator (Y-axis) with the MRLs proposed by EFSA or JMPR experts as well as with the MRLs produced by the NAFTA Calculator (X-axis). Both axes are represented using a logarithmic scale. The points on the blue line correspond to datasets for which the OECD Calculator yields a MRL-estimate that is equal to the MRL proposed by experts or produced by the NAFTA Calculator. Points above (below) the line correspond to datasets for which the OECD Calculator yields a MRL-estimate that is higher (lower) than the MRL proposed by experts or produced by the NAFTA Calculator. A point may represent several datasets.

Figure 10. Comparison between the MRLs produced by the OECD Calculator and the MRLs proposed by EFSA experts

(EFSA datasets).

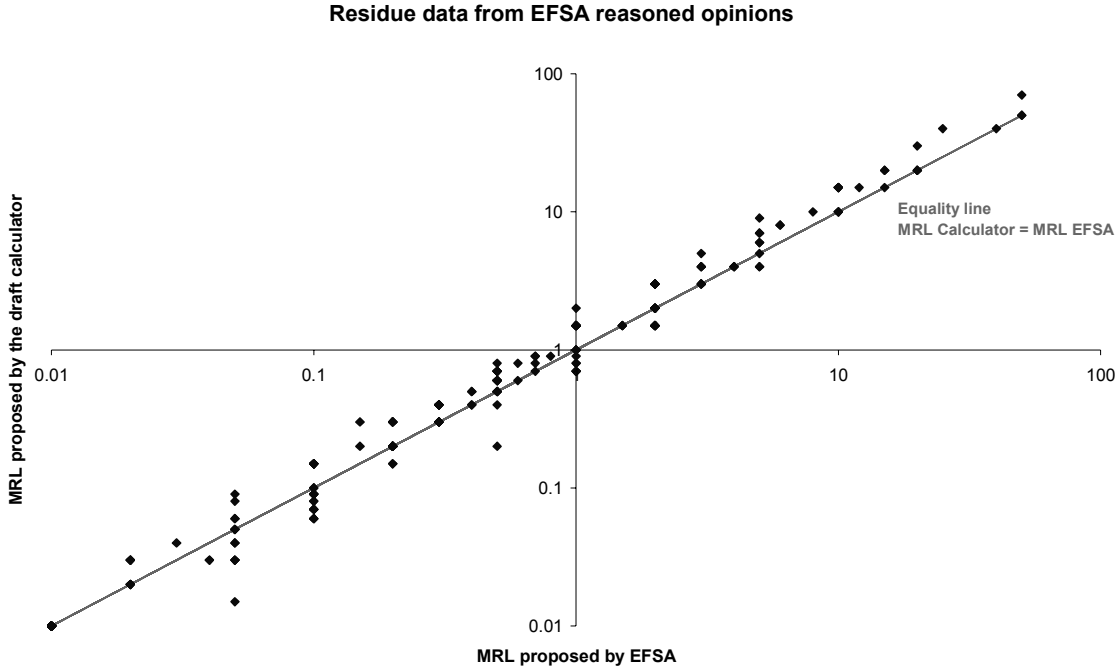


Figure 11. Comparison between the MRLs produced by the OECD Calculator and the MRLs produced by the JMPR experts (JMPR datasets).

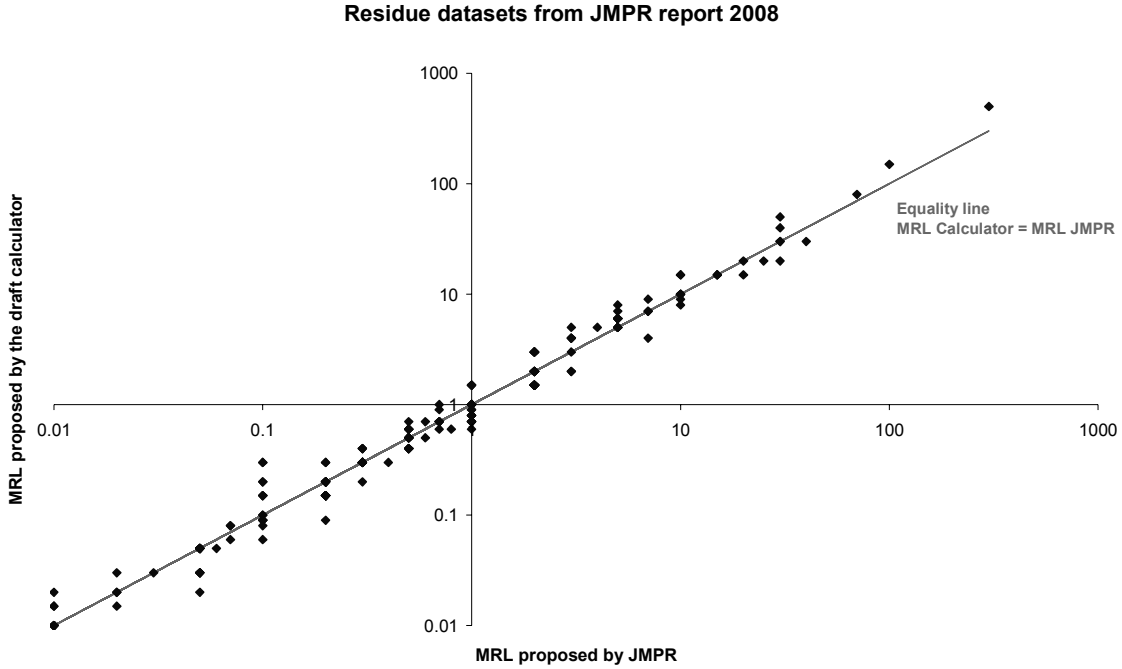
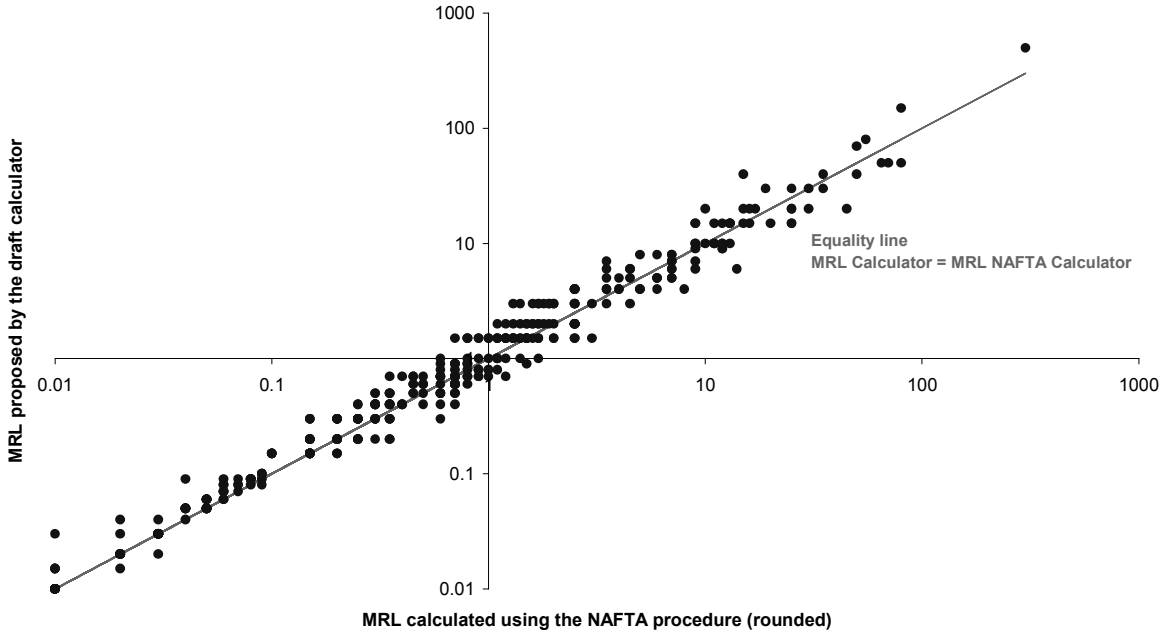


Figure 12. Comparison between the MRLs produced by the OECD Calculator and the MRLs produced by the NAFTA Calculator (combined EFSA and JMPR datasets).



REFERENCES

OECD Guidelines for the Testing of Chemicals. Test No. 509: Crop Field Trial. Adopted 7th September 2009.

Environmental statistics and data analysis. Wayne R. Ott. CRC Press, 1995.

Guidelines for the generation of data concerning residues as provided in Annex II, part A, section 6 and Annex III, part A, section 8 of Directive 91/414/EEC concerning the placing of plant protection products on the market. Commission of the European Communities.

Appendix D. *Comparability, extrapolation, group tolerances and data requirements.* Directorate General for Health and Consumer Protection (SANCO E.1). Doc. 7525/VI/95, revision 7, 12th June 2001.

Appendix I. *Calculation of Maximum Residue Levels and Safety Intervals e.g. Pre-harvest Intervals.* Directorate General for Agriculture (VI B II-1). Doc. 7039/VI/95 EN, 22nd July 1997.

Does calculation of the 95th percentile of microbiological results offer any advantage over percentage exceedence in determining compliance with bathing water quality standards? P. R. Hunter. Letters in applied Microbiology 2002, vol. 34, 283-286.

Statistical Methods in Water Resources (chapter 13). D. R. Helsel and R. M. Hirsch. United States Geological Survey. http://water.usgs.gov/pubs/twri/twri4a3/

Nondetects and Data Analysis. Dennis R. Helsel. Wiley-Interscience 2005.

Maximum Residue Levels: Fact or Fiction? Kieran Hyder, Kim Z. Travis, Zoe K. Welsh, and Ian Pate. Human and Ecological Risk Assessment: Vol. 9, No. 3, pp. 721-740 (2003).

Guidance for setting pesticide maximum residue limits based on field trial data. US EPA Office of Pesticide Programs and Health Canada PMRA, 28th September 2005.

http://www.epa.gov/oppfead1/international/naftatwg/index.html

Statistical Basis of the NAFTA Method for Calculating Pesticide Maximum Residue Limits from Field Trial Data. US EPA Office of Pesticide Programs and Health Canada PMRA.

APPENDIX A: MEAN + K*SD APPROACHES

As an alternative to introducing a "floor" method to reduce underestimations of the 95th percentile for very small datasets, a number of other procedures were investigated. Most of these procedures adjusted the number of standard deviations used in the calculation according to dataset size.

By the "Mean + k*SD" method we mean that k is an integer factor[7] that depends on the number of data points n, so we have k = k(n). If the value of "Mean + k(n)*SD" is less than the highest residue (HR), then the HR is taken as a MRL proposal. So the MRL proposal is

Maximum (mean + k(n) * SD, HR).

Several criteria were suggested to calculate the k factors, i.e. the functional dependency of k on the number of data set size n. Below we report the results of just one of them; the other criteria produced similar results.

Let the target criteria be that MRL proposals are 1) higher than p95, in line with the EU and NAFTA recommendations, and 2) lower than F*p95 (F is an integer factor) with a high level of confidence q. The default level of confidence was set to q = 95 %, in line with the EU and NAFTA recommendations. However, for small datasets, simultaneous satisfaction of requirements 1) and 2) turn out to be impossible. In such cases the level of confidence was decreased as a function of dataset size. Factor F was set to 4 to provide a) not too high MRL estimates and b) reasonable levels of confidence for small datasets.

To derive the dependencies k(n) and q(n), 100,000 synthetic datasets sampled from the lognormal distribution with the CV of 1.0 were considered. For each dataset in each data size class, a MRL proposal was calculated. A k value was accepted for a given number of data points if upper and lower confidence limits of MRL proposals were below 4*p95 and above 1.0*p95, respectively (requirements 1) and 2)). If necessary, the confidence level q was decreased from 95 %.

Calculated values of k varied from 9 for 3 data points to 4 for 39 or more data points (Fig. 12). The level of confidence was in the worst case (3 data points) 60 %, while the level of 95 % was reached for 22 data points (Fig. 13). In the worst case (3 data points) most of the calculated MRL proposals (95 %) lay between 0.4*p95 and 9.0*p95 (Fig. 14), while the MRL/HR ratio varied between 1.9 and 5.1 (Fig. 15). On the other hand, if MRL proposals were considered at calculated levels of confidence, they would be, by the definition of the procedure, within required upper and lower boundaries for all dataset sizes (Fig. 14, grey labels). The failure rate, i.e. the chance to get a MRL below the p95, was at maximum 19.1 % for n = 3 data points (Fig. 16), and fast decreased to a level of about 2–3% with an increase of n.

[7] We use integer values for simplicity; similar but smoother results are obtained if k is varied continuously.

Figure 13. Calculated k values depending on the number of data points.

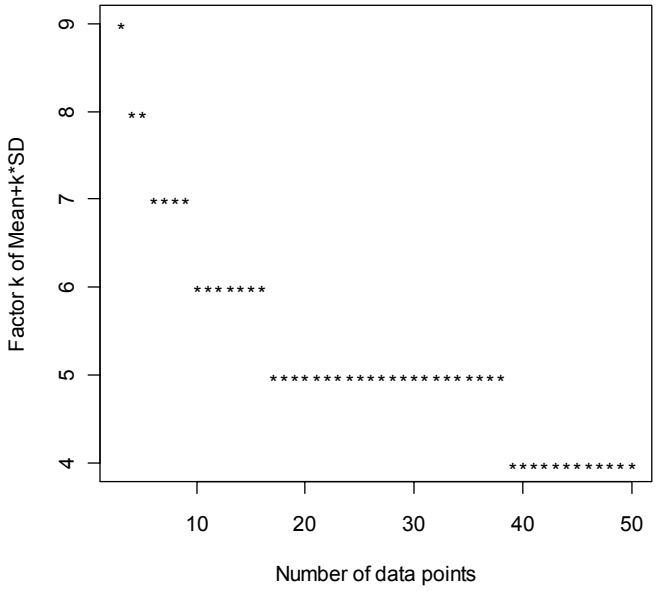


Figure 14. Calculated levels of confidence depending on the number of data points.

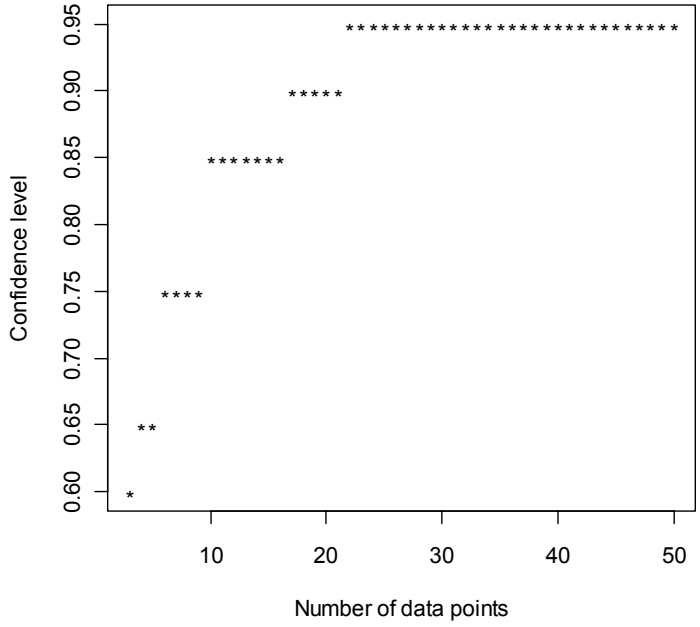


Figure 15. 95% probable ranges for the MRL-over-p95 ratio depending on the number of data points. Red (+) and green (*) labels show the upper and lower boundaries, respectively. Blue lines show the levels of 1*p95 and 4*p95. Black labels (o) show values calculated using 1*p95 and 4*p95 as lower and upper limits for MRL proposals.

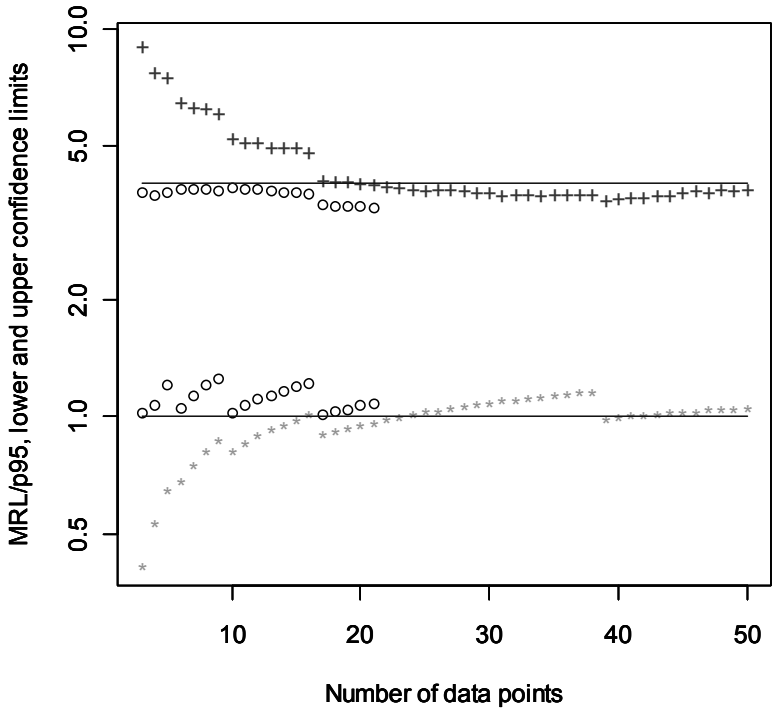


Figure 16. 95% probable range for the MRL-over-HR ratio. Red (+) and green (*) labels show the upper and lower boundaries, respectively.

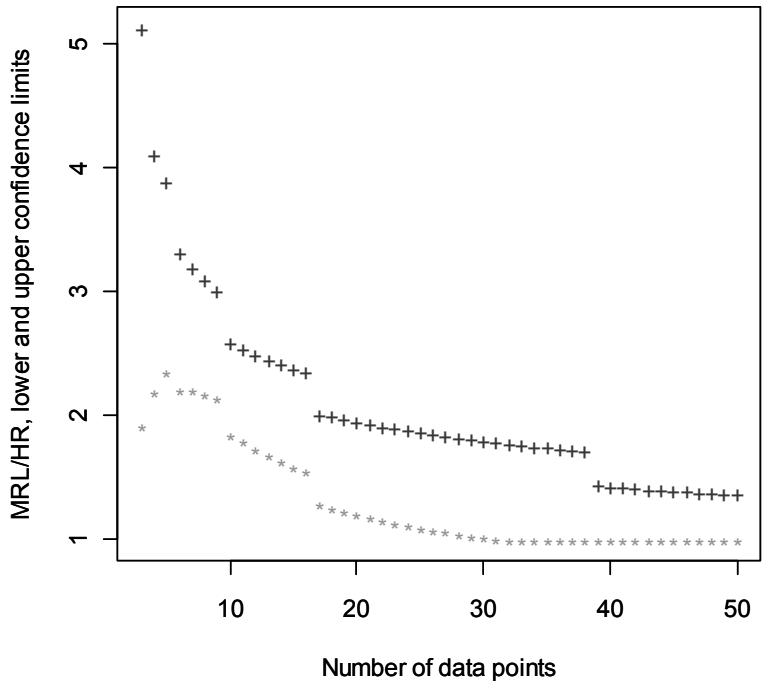
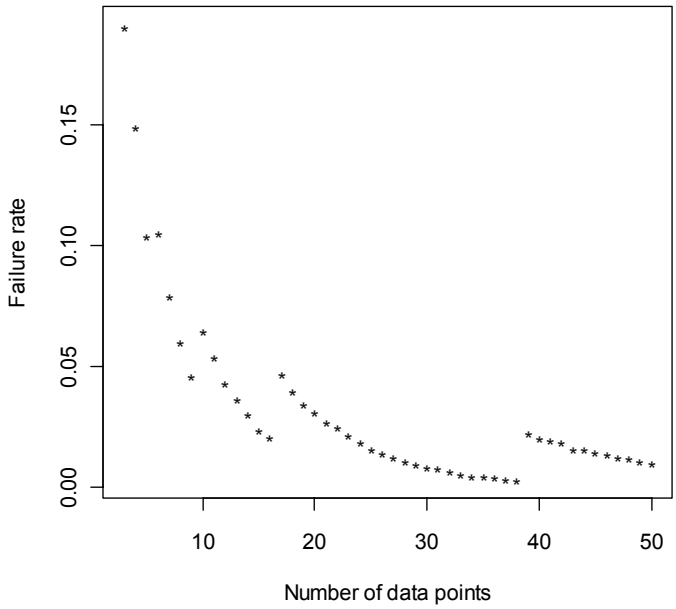


Figure 17. "Failure rate", i.e. the fraction of datasets for which the proposed MRL is below the p95, versus the number of data points.



The "Mean + k*SD" method was tested using real residue data from EFSA and JMPR evaluations. The test results are summarised in table 4 and table 5.

Table 4. Tests with the "Mean + k*SD" method using EFSA residue data.

EFSA datasets	Rounded MRL / Unrounded MRL	Rounded MRL / HR	Rounded MRL / EFSA MRL	Rounded MRL / NAFTA calc. MRL
Min (overall)	0.91	1.00	0.40	0.67
Max (overall)	10.31	10.31	3.00	3.75
Mean (overall)	1.19	2.86	1.59	1.53
Mean (n ≤ 4)	1.33	3.45	1.51	1.55
Mean (4 < n ≤ 8)	1.16	2.94	1.65	1.59
Mean (8 < n ≤ 16)	1.12	2.26	1.63	1.42
Mean (n > 16)	1.06	1.43	0.97	1.14
Mean (0% censored)	1.12	2.94	1.68	1.58
Mean (100% censored)	1.03	1.03	0.91	1.00

Table 5. Tests with the "Mean + k*SD" method using JMPR residue data.

JMPR datasets	Rounded MRL / Unrounded MRL	Rounded MRL / HR	Rounded MRL / JMPR MRL	Rounded MRL / NAFTA calc. MRL
Min (overall)	0.95	1.00	0.50	0.63
Max (overall)	1.42	5.00	5.00	2.86
Mean (overall)	1.09	2.37	1.38	1.47
Mean (n ≤ 4)	1.08	3.12	1.74	1.70
Mean (4 < n ≤ 8)	1.09	2.82	1.56	1.56
Mean (8 < n ≤ 16)	1.07	2.08	1.26	1.41
Mean (n > 16)	1.12	1.65	1.06	1.32
Mean (0% censored)	1.09	2.62	1.41	1.51
Mean (100% censored)	1.00	1.00	1.04	1.00

On average the rounding procedure tends to increase the MRLs by about 10—20%. With some EFSA datasets, a more than 10-fold increase due to rounding was observed. This is because the LOQ for these datasets was 0.00097 mg/kg and the unrounded MRLs were far below the lowest MRL class of 0.01 mg/kg.

The mean ratio between the rounded MRL and the HR was 2.9 and 2.4 for the EFSA and JMPR dataset collections, respectively. This ratio tends to decrease when the size of the dataset increases. A more than 10-fold ratio was observed for some of the EFSA datasets that were analysed with an LOQ of 0.00097 mg/kg and which contained no data higher than the LOQ.

On average the MRL estimates yielded by the OECD Calculator exceed the MRLs proposed by EFSA and JMPR experts by 59% and 39%, respectively. However, larger deviations are observed for some individual datasets. The ratio between the MRL estimates produced by the OECD Calculator and the MRLs proposed by experts ranges between 0.40 and 3.0 for the EFSA datasets and between 0.50 and 5.0 for the JMPR datasets.

On average the MRL estimates yielded by the OECD Calculator also tend to exceed the MRLs yielded by the NAFTA Calculator. The average difference is 53% for the EFSA datasets and 47% for the JMPR datasets. Again, larger deviations are observed for some individual datasets. The ratio between the MRL estimates produced by the OECD Calculator and the MRLs produced by the NAFTA Calculator ranges between 0.67 and 3.8 for the EFSA datasets and between 0.63 and 2.9 for the JMPR datasets.

The following graphs allow us to compare the MRLs produced by the "Mean + k*SD" method (Y-axis) with the MRLs proposed by EFSA or JMPR experts as well as with the MRLs produced by the NAFTA Calculator (X-axis). Most points are located above the blue line, which indicates that the tested "Mean + k*SD" method tends to produce MRLs that are both higher than the MRLs recently proposed by EFSA and JMPR experts and higher than the MRLs produced by the NAFTA Calculator.

Figure 18. Comparison between the MRLs produced by the "Mean + k*SD" method and the MRLs proposed by EFSA experts

(EFSA datasets).

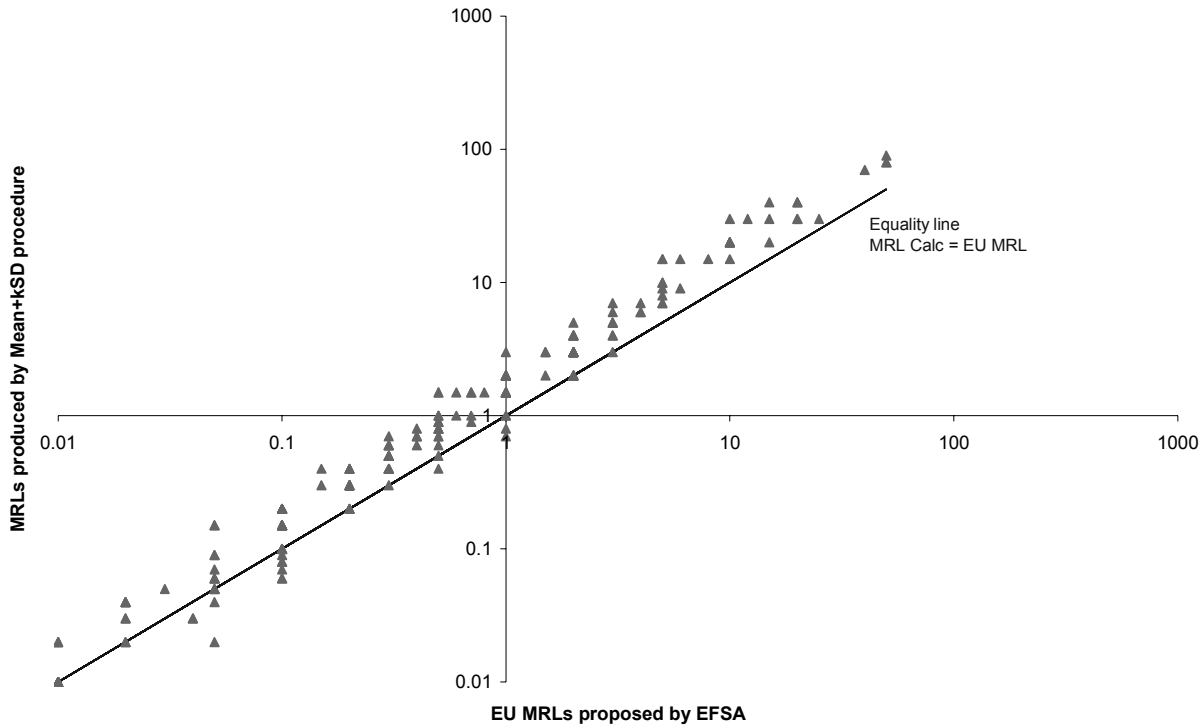


Figure 19. Comparison between the MRLs produced by the "Mean + k*SD" method and the MRLs proposed by JMPR experts.

(JMPR datasets).

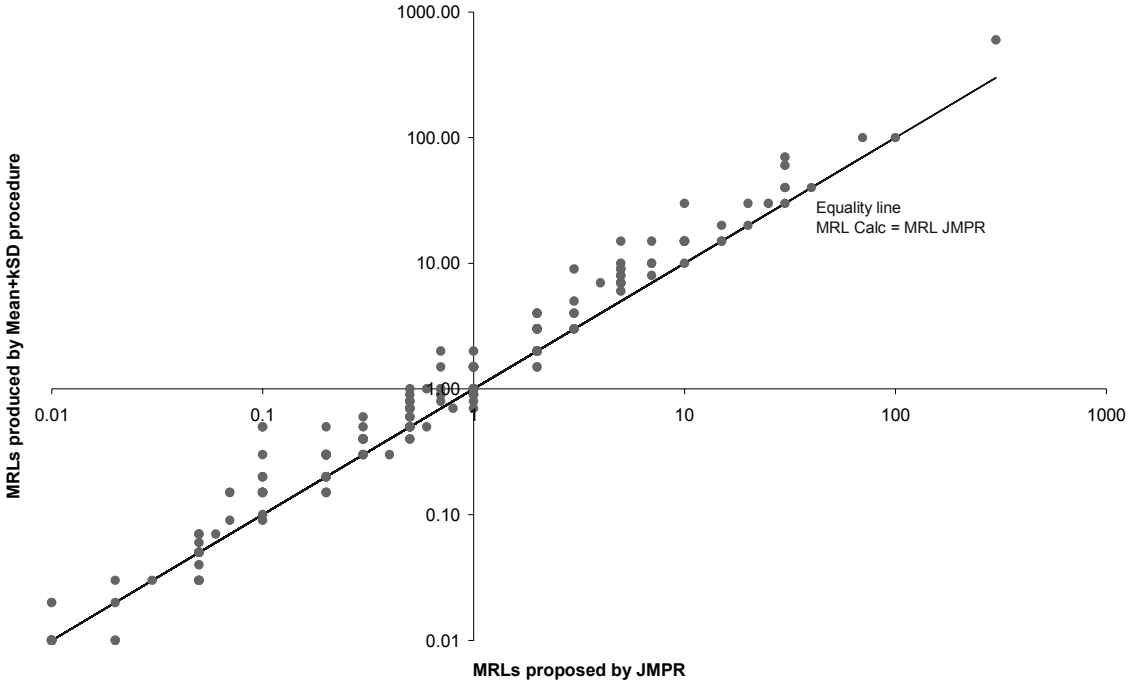
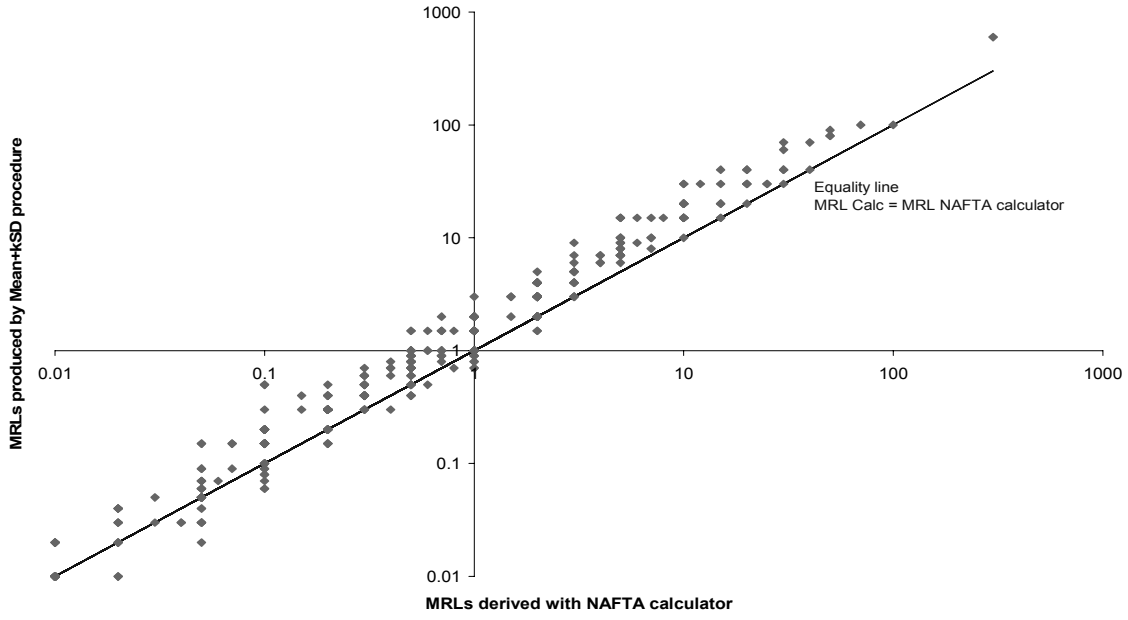


Figure 20. Comparison between the MRLs produced by the "Mean + k*SD" method and the MRLs produced by the NAFTA Calculator (combined EFSA and JMPR datasets).



In general, this method is more conservative, not only more than the one suggested by the MRL calculation group, but also more than any other MRL calculation procedures ever used before.

Conclusions

All the procedures included in this section are scientifically sound, being based in well established statistical procedures. They differ on their degree of conservatism and their performance with small datasets.

The suggested base method, "Mean + 4*SD", was always expected to be more robust and outperform distributional methods for very small datasets. That is why a similar method to this one was already introduced as part of the NAFTA Calculator. But it was a pleasant surprise for the calculator group members to find out that it also outperforms or at least, performs similarly to the distributional methods for datasets of 20 to 30 points.

Our efforts to reduce the underestimation of the 95th percentile for very small datasets by making the number of standard deviations dependent on the dataset size produced much more conservative methods and increased overestimations. The addition of a floor method was a pragmatic decision born from this experimentation and it is oriented to correct the worst underestimations without increasing the overestimations.

In relation to the overestimations, the issue of potential "outliers" was discussed many times by the group. It was decided that it is very difficult to classify a certain high value as an outlier for small datasets, since there is not enough information in them to determine the "trend" of the dataset. For larger datasets, which have been shown to clearly follow a certain distribution (normal, lognormal, Weibull, etc), an appropriate outlier test to that distribution may be performed; e.g., the Dixon test on the residues directly for the normal distribution, or in the logarithms of the residues for the lognormal distribution. This may highlight certain residue measurements that deserve further investigation. But the group would not recommend ignoring these data points exclusively based on statistical tests.

APPENDIX B: DISTRIBUTIONAL VERSUS NON-DISTRIBUTIONAL APPROACHES

The original design of the calculator was based on the expectation that distributional approaches (fitting to normal, lognormal and Weibull distributions) would be more accurate than non-distributional approaches for large datasets. However, the calculator work group has found that the "Mean + 4*SD" method outperformed the distributional approaches, not just for small datasets, but also for datasets as large as 20 or 30 points.

To compare the performance of the proposed method with the previously used distributional method, 10,000 datasets were sampled from the lognormal, normal and Weibull distributions with CV = 1.0 for each dataset size from 3 to 30.[8] For each distribution and for each number of data points, the 95% probable range[9] of the ratio of the calculated MRL to the true p95 was calculated. Negative values sampled from the normal distribution were replaced with 0.001.

Figures 21—23 demonstrate that in general the proposed method outperformed the distributional ones for dataset sizes from 16 to 30 (for smaller datasets the distributional method was not applied). Indeed, for lognormal and Weibull datasets (Fig. 20 and 22, respectively), the probable range was narrower, and both the upper and lower boundaries of the MRL-over-p95 ratio were closer to 1.0 compared to the distributional method, i.e. the 95th percentile of the residue distribution was estimated more accurately. For normal datasets the distributional method provided the MRL-over-p95 closer to 1.0 and narrower probable range (Fig. 21). On the other hand, the proposed method performed more conservatively; the entire probable range was above the 95th percentile for datasets of 16 and more points. The distributional method, in contrast, underestimated the 95th percentile of the underlying normal distribution even for medium size datasets. Similar results were found for CV ranging from 0.5 to 1.5.

[8] Lognormal distribution with the mean of logs = 1.0 and SD of logs = 0.83, normal distribution with the mean = 1.0 and SD = 1.0 and Weibull distribution with the shape parameter = 1.0 and scale parameter = 1.0 were considered.

[9] The interval between 2.5% lowest value and the 97.5% highest value of the ratio computed from all the datasets.

Figure 21. Lognormal datasets: 95 % probable ranges for the MRL-over-p95 ratio depending on the number of data points. Red and green labels show the upper and lower boundaries, respectively, calculated by the proposed method. Blue labels show values calculated by the distributional method.

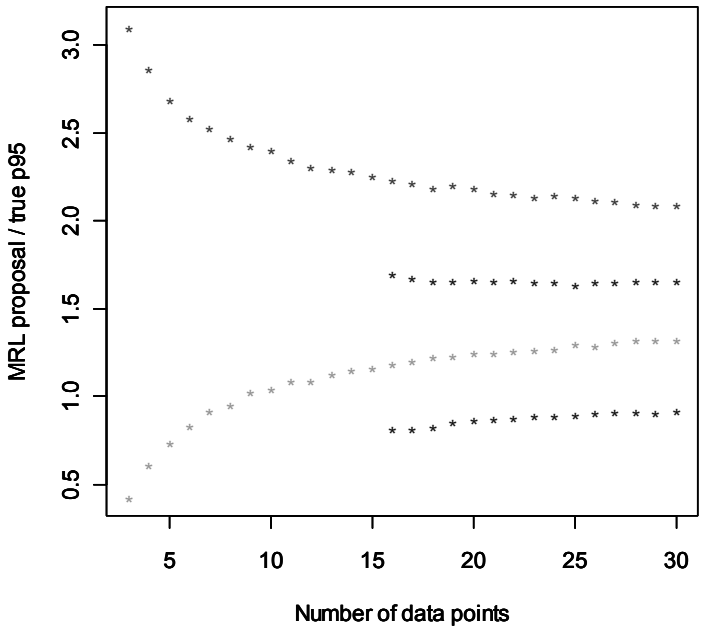


Figure 22. Normal datasets: 95 % probable ranges for the MRL-over-p95 ratio depending on the number of data points. Red and green labels show the upper and lower boundaries, respectively, calculated by the proposed method. Blue labels show values calculated by the distributional method.

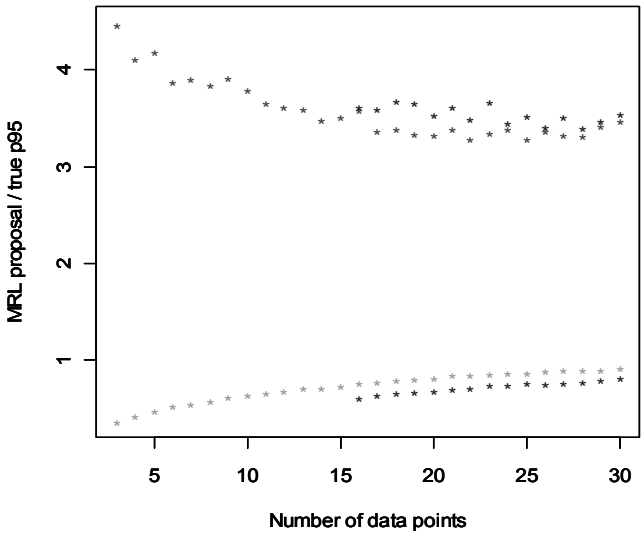
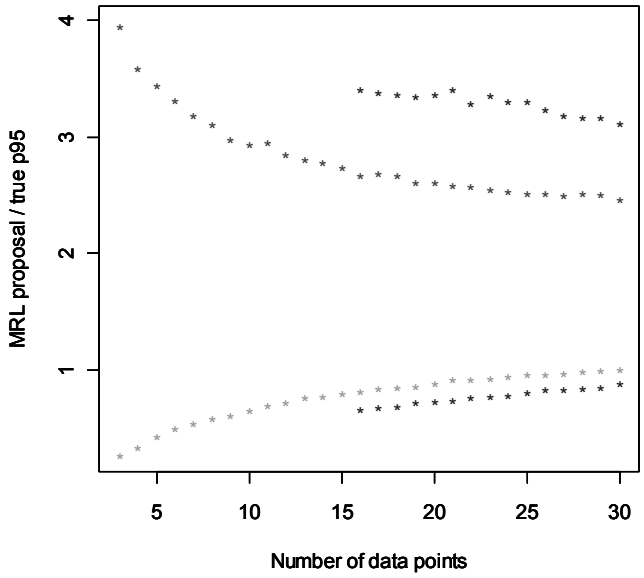


Figure 23. Weibull datasets: 95 % probable ranges for the MRL-over-p95 ratio depending on the number of data points. Red and green labels show the upper and lower boundaries, respectively, calculated by the proposed method. Blue labels show values calculated by the distributional method.



APPENDIX C: STATISTICAL REASONING FOR FULLY CENSORED DATASETS

At its meeting in August 2010, the OECD RCEG proposed that the MRL for fully censored datasets be set at the level of the highest LOQ present in the dataset, thereby deciding not to adopt the statistically-based recommendation made in this appendix.

In the situation where all n data are below a single LOQ, the "Mean + 4*SD" could mathematically be anywhere between 0 and roughly 2.5*LOQ (perhaps as much as 3*LOQ at small sample sizes). Making assumptions/requirements about the CV does not provide any further restriction. Consequently, we needed a different approach. The only solid information we had was that HR<LOQ and so we sought a sensible procedure based on the HR and then substituted LOQ for HR. In what follows, the guiding principle is that we wanted the MRL proposal to have at least a 50% chance of exceeding p95 (the 95th percentile of residues).

In general, for any continuous distribution, the probability that HR is less than p95 is $(0.95)^n$. For $n \geq 14$, $(0.95)^n < 0.5$ and so the HR itself has at least 50% chance of exceeding the p95. However, for n=3, $0.95^3 = 0.86$ and so there is only a 14% chance that the HR exceeds the p95. It did not seem reasonable to set MRL=HR for small sample sizes such as n=3 on a purely statistical basis.

But it did make sense to set the MRL at some multiple of HR, for example 2*HR. The probability that 2*HR is less than 95th percentile depends on the distribution. Figure 24 shows that 2*HR has at least a 50% chance of exceeding p95 for log-normal and Weibull distributions with CV<1.5 and $n \geq 4$. This also holds for larger CV for n>=5 and for CV<1 for n=3. It therefore seemed reasonable to set MRL=2*HR for $n \geq 4$ and not unreasonable to extend this to n=3 in the interest of simplicity. We can also see that the chance of exceeding the p95 is substantially greater than 50% as n increases. We therefore considered something between HR and 2*HR for some sample sizes.

Figure 25 shows that 1.5*HR has at least 50% chance of exceeding 95th percentile for log-normal and Weibull distributions with CV<2 and $n \geq 8$. It therefore seemed reasonable to set MRL=1.5xHR for $n \geq 8$.

Finally, we chose to switch from 1.5*HR to HR at n=16 rather than n=14 as the point where procedure changes in order to keep some consistency in terms of where changes of procedure occur in the calculator. It has the added advantage of restoring a little of the conservatism demonstrated by methods used when there is at least one measurement above the LOQ.

Combining the foregoing arguments leads to the calculation proposed in the main body of the white paper. A final comment is that our reason for using "Mean + 4*SD" and other mean based methods where possible is because they have lower sampling variability than procedures based on the HR. The calculations above demonstrate a 50% chance of exceeding the 95th percentile in many situations but do not show the variability of the ratio of MRL to 95th percentile.

Figure 24. Performance of 2*HR relative to p95 for Weibull and log-normal distributions for data set sizes from 3 to 7 and a wide range of coefficients of variation.

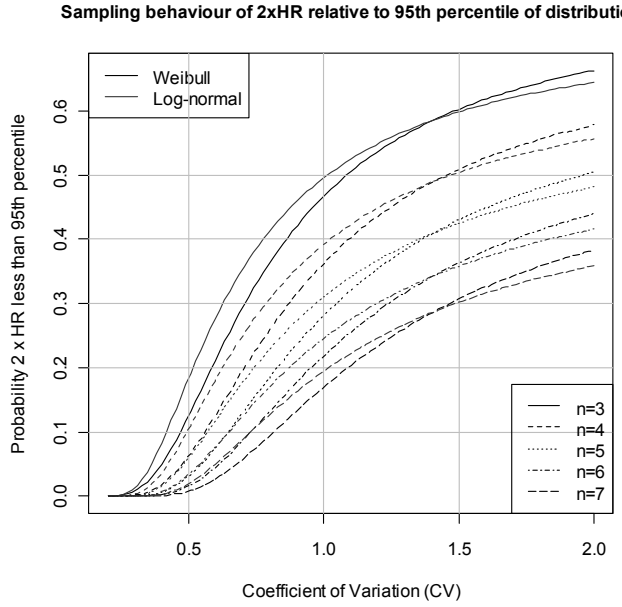
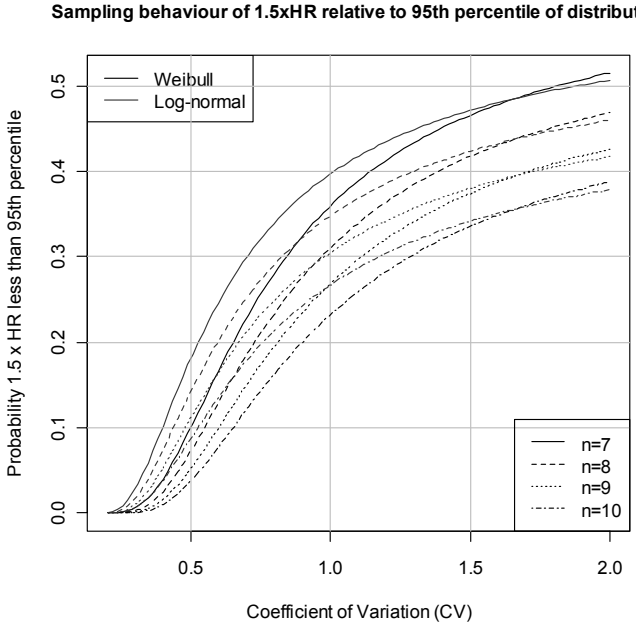


Figure 25. Performance of 1.5*HR relative to p95 for Weibull and log-normal distributions for data set sizes from 3 to 7 and a wide range of coefficients of variation.



Where there is more than one LOQ involved, the calculator follows the same procedure based on the maximum LOQ. However, it should be recognized that there may be grounds for setting a lower MRL in such situations depending on the balance of the number of measurements below each of the different LOQs.

It should also be recognized that in any fully censored situation there may be other considerations which would lead to setting a lower MRL than that proposed by the calculator group. For example, there might be evidence, independent of the residue trials, suggesting that residues were in fact expected to fall well below the LOQ.

APPENDIX D: JUSTIFICATION FOR USE OF AVERAGE VALUES FROM FIELD TRIAL REPLICATES WHEN CALCULATING MAXIMUM RESIDUE LEVELS

The OECD MRL Calculator is based on the general equation "Mean + 4*SD" method, where SD is the standard deviation of the pesticide residue values obtained from independent supervised field trial measurements. The same principles which guide the current OECD and FAO crop field trial practices provided the framework for the development of the OECD Calculator methodology.

According to both FAO and OECD guidance, residue data in crops are generated from multiple field trials, established in separate sites or trials. "Representative" sample(s), which are composite samples, are collected from each field trial plot. Composite samples are created by combining multiple sampling portions from random or systematically determined sampling points within the plot. According to the FAO manual Guidelines on Producing Pesticide Residues Data from Supervised Trials, "The samples must be representative to enable the analytical result to be applied to the entire experimental unit." It is understood, therefore, that the goal of collecting composite samples from supervised crop field trials is to estimate the mean or "true" single field trial residue level at a specific time point resulting from the application of a pesticide product.

Replicate Measurements – a basic analytical principle

The collection and analysis of replicate measurements is a basic principle of analytical chemistry (see [10]). The analytical principle involved in replicate testing is that the average of multiple replicates leads to a better estimate of the true mean of the underlying sample population. If the variability between replicates is considered to be indeterminate in nature (influenced by random uncontrollable factors which result in an equal probability of being greater or less than the "true" value), than it is true that the average value is likely to be more reliable than any individual replicate value.

This principle is adhered to in today's regulatory environment when replicate measurements are conducted in the laboratory (multiple analysis of the same field sample). In this case, it is understood that the best answer (the answer closest to the "true" residue value of the field sample) is the average of the laboratory measurements. It follows that when attempting to determine the "true" residue value of a crop field trial, the best answer is the average of the field trial replicates.

The practice of considering only the maximum field trial replicate when multiple replicates have been collected violates these basic principles of analytical chemistry. Instead of estimating the best "true" field trial residue, considering only the maximum residue introduces a systematic bias - a determinate error in the positive direction which skews the results. This might be considered a conservative strategy intended to minimize underestimation of the MRL, but the current basis of the OECD MRL Calculator make this completely unnecessary. For most data sets, the Mean + 4SD calculation results in a theoretical percentile that is higher than 95th percentile. Testing has shown that the MRL proposals originating from the current version of the OECD MRL Calculator tend to be higher than the MRLs established by the European authorities and JMPR using the same datasets.

Variability

Averaging replicate measurements at any level (analytical, sampling, etc) can reduce variability at that level. If the measurements are not skewed (and therefore they are symmetrically randomly distributed around the mean), the average of the replicate measurements will give an estimate of the mean, or true value, of the residue at that level. Variability of residue data is inherent at every level of the process. At the lowest level, there is analytical instrument variability. At the next level (sample level), analytical method variability and subsampling variability (combined with instrumental variability) contribute to variability between sample analyses. Although current guidelines do not require replicate injections or replicate analyses, if the laboratory does conduct replicate injections or replicate analyses, the replicate values would be averaged, as appropriate, thereby giving the best estimate of the true residue value and reducing the variability from these sources.

At the trial level, sampling variability and sample compositing variability contribute to variability between replicates (intra-site or intra-plot variability). Again, if replicates are taken at this level (within one plot), the replicate values should be averaged, reducing the variability from these sources.

Finally at the highest level, there are the environmental factors, investigator practices, equipment differences (and other known and unknown factors) which could lead to variability between trials or sites. The sum of these inter-site variations is significantly larger than the variations at the lower levels, and it is this inter-site variability that is the important residue data variable from which the MRL is intended to be derived.

The variability between single field trial samples (instead of replicates), includes the variability from all levels of the process, including intra-site variability. This intra-site variability is significantly smaller than the inter-site variability, and has little influence on the overall variability of residue values and, therefore, the MRL proposal.

The variability between replicate averages (average values from replicate field samples), also includes the variability from all levels of the process; however, the intra-site variability is reduced by an amount relative to the number of replicates that are averaged. As in the case for single replicates, the MRL proposal is driven predominately by the inter-site variability.

If only the maximum replicates are considered instead of the average value of replicates, the variability between the replicates includes both the intra-site variability (just as the case for single samples) and inter-site variability; however, **the residue values no longer give the best estimate of the true field residue values. In fact, the residue values are biased in a positive direction and may lead to significant over-estimation of some MRLs.**

Testing Results using Replicate Values

101 sets of field trial data from one recent JMPR submission and two recent NAFTA submissions were investigated using the July 2010 version of the OECD MRL Calculator. Each trial in this collection of residue data was represented by two field replicates. For each set of trial data, MRLs were calculated using:

- the average of each replicate pair;
- the maximum of each replicate pair;

- alternating replicates (rep A from first pair, rep B from second pair, rep A from third pair…).[10]

The MRLs from the averaged replicates were compared to the MRLs from the maximum replicates and the MRLs from the maximum and average replicates were compared to the MRLs from alternating replicates. Standard deviations and means were also compared for these data sets.

The results show that the rounded MRLs generated by using the maximum replicate values versus the average replicate values are either the same (60% of the time) or higher (40% of the time). This is due partly to the fact described above that the standard deviations are slightly lower than they would be if only single replicates were collected due to reduction of the intra-site variability, but also to the fact that the calculated means of the residue data sets comprised of the maximum replicates are almost always higher (94% of the time) than the means calculated from the data sets comprised of the average replicate values.

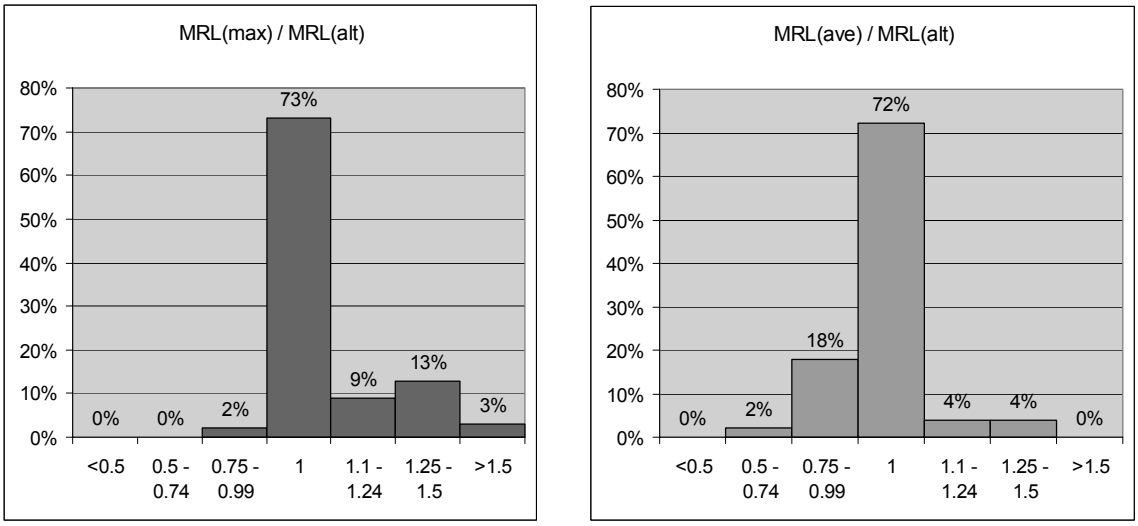
When MRLs generated using average replicates values were compared to the MRLs generated using alternating replicate values, the MRLs from the average replicate values were usually the same, but fluctuated both lower and higher (20% lower, 72% the same, 8% higher). As expected, the standard deviations were slightly lower for the sets with averaged replicates (based on reduced intra-trial variability), but the means tended equally lower and higher.

When MRLs generated using maximum replicates values were compared to the MRLs generated using alternating replicate values, the MRLs from the maximum replicate values were usually the same, but sometimes tended to be higher (2% lower, 73% the same, 25% higher - - 7% were over 1.5x higher). This was due to solely the fact that the means of the residue data sets were generally higher (86% of the time).

It is these results which demonstrate that considering only the maximum replicate values violates the basic principles of analytical chemistry - that the practice of collecting and analyzing replicates should lead to a higher MRL than the practice of collecting only single samples. Consider the argument that the more replicates that are collected and analyzed, and thus, the more data that are generated, the higher the MRL will be, based on this practice.

[10] The alternating replicates were tested to get an idea of what the MRL would be if only one replicate had been collected from each site.

Figure 26. The Figure on the left shows a comparison of the MRLs derived from data sets using the maximum replicate values to MRLs derived from data sets using alternating replicates (ratio MRL(max) to MRL(alt)). The Figure on the right shows a comparison of the MRLs derived from data sets using the average replicate values to MRLs derived from data sets using alternating replicates (ratio MRL(ave) to MRL(alt)).



Response to 2007 JMPR Report

In the 2007 JMPR report (see [11]) the use of the terms standard deviation (SD) and standard error (SE) is confusing. A description of the two terms, as they might relate to field trial residue data is included below (see also [12]).

Standard deviation

Standard deviation can be thought of as the "average distance" of each individual measurement, in this case "each individual residue", from the mean measurement (i.e. mean residue). Generally, a good estimator can be found for the standard deviation that is not always high nor is it always low for small sample sizes (i.e. an unbiased estimator of the standard deviation), which means that collecting more samples does not necessarily mean that the estimate of the standard deviation will decrease. Additionally, the more samples you have, the more accurate this estimate becomes. However the estimate of the standard deviation never goes to zero, since the "average distance" between the individual residues and the mean residue is never zero.[11]

Standard error

The standard error is a measure of the precision of an estimator. If not further specified, the standard error usually refers to the standard error of the mean. (However, any population estimate such as the mean,

[11] The actual "average distance" from the mean is always zero, which is why the difference or "distance" between each observation and the mean is squared (which effectively confines the calculation to absolute distances) when calculating the standard deviation.

standard deviation, or 95th percentile can have a standard error, that is, a measure of the precision of that estimate). For a normal distribution, the standard error is the sample standard deviation (i.e. the estimate of the standard deviation based on the sample) divided by the square root of n (the number of samples). Unlike the standard deviation, the standard error does approach zero as the sample size increases. This is because the estimate of the mean becomes more precise as the sample size increases. You can think of the standard error of the mean as the "average distance" between several estimates of the mean and the actual (true) mean. Imagine having 10 field trials (with one residue from each trial) and calculating the mean: this is an estimate of the true mean. Now, imagine that this is repeated 50 times: 50 sets of 10 field trial residues would result in 50 estimates of the true mean. The standard error of the mean measures the variability between the 50 estimates of the true mean. Now imagine that the number of field trials in each set was increased to 100. Each of the 50 estimates (each based on 100 samples) of the true mean would be more precise or "closer" to the true mean and there would be less variability between these 50 estimates of the mean. Eventually, if enough field trial samples were collected in each of the 50 sets (e.g. 10,000 field trial residues), the variability between the 50 estimates of the mean would be practically zero. On the other hand, the variability (and thus the standard deviation) between even 10,000 individual field trial residues would not be zero, but would only more accurately measure the average distance between these individual residues and the mean residue.

In summary, the difference between the two is that the standard deviation is the "average distance" between the individual residues and the mean residue, while the standard error is a measure of the precision of a population parameter, usually the mean.

The 2007 JMPR report states, "Where the average residue measured in replicate random samples taken from one field would be used as a single residue value the true distribution of the residues would be apparently reduced proportional to the square root of the number of replicate field samples.

According to the sampling theory the standard deviation (also called the ‘standard error’) of mean residue in n samples taken from the populations of “i” samples is:

$$S_n = S_i / \text{sqrt}(n).$$

Consequently if we use the average of two randomly selected replicate field samples taken from a plot, we reduce the standard deviation of the residues by 1.41."

The report seems to suggest that the standard error is reduced when the sample size is increased. This is true, but the MRL calculation used in the OECD Calculator (i.e. Mean + 4SD) does not make use of the standard error. Furthermore, when an unbiased estimate of the standard deviation is used, the standard deviation is not necessarily reduced as the sample size increases.

Lastly, it is important to note that if we do use the average of two field replicates (as opposed to the maximum), the "n" in the denominator (of the standard deviation calculation) would represent the number of field trial locations, not the number of replicates. For example, if two samples were taken at each of five field trial locations, then the "sample size" used in the standard deviation calculation would be 5, not 10. In other words, the denominator in the standard deviation calculation based on the average field trial residues would be the same as that proposed by JMPR when using maximum field trial residues.

REFERENCES TO THE ANNEXES

Fundamentals of Analytical Chemistry, 2nd edition. D. A. Skoog and D. M. West.

Pesticide residues in food 2007. Joint FAO/WHO Meeting on Pesticides Residues. Report 2007.

Altman, D.G. and Bland, J.M. Standard Deviations and Standard Errors. BMJ 331:903, 2005.