CODEX ALIMENTARIUS COMMISSION

Food and Agriculture Organization of the United Nations

World Health Organization

Viale delle Terme di Caracalla, 00153 Rome, Italy - Tel: (+39) 06 57051 - E-mail: codex@fao.org - www.codexalimentarius.org

Agenda Item 17 **CX/CF 21/14/15**

April 2021

ORIGINAL LANGUAGE ONLY

JOINT FAO/WHO FOOD STANDARDS PROGRAMME

CODEX COMMITTEE ON CONTAMINANTS IN FOODS

14th Session

(virtual)
3-7 and 13 May 2021

GUIDANCE ON DATA ANALYSIS FOR DEVELOPMENT OF MAXIMUM LEVELS AND FOR IMPROVED DATA COLLECTION

(Prepared by the European Union as
Chair of the Electronic Working Group)

BACKGROUND

1. At its 12th session, CCCF considered the proposal of the JECFA Secretariat to develop a general guidance on data analysis for ML development as it was observed that different approaches were taken by the EWGs. These differences concerned for example the handling of occurrence data without information on LOQ. A general guidance would help future EWGs to take consistent approaches for data analysis. CCCF agreed to establish an EWG chaired by EU, co-chaired by the United States of America, the Netherlands and Japan, working in English, to prepare a discussion paper[1]

2. At its 13th session, the EU as chair of the EWG, informed the CCCF that it has not been possible to prepare in time a discussion paper for consideration by the established EWG. Therefore, a paper prepared by the EU as Chair of the EWG containing a non-exhaustive list of topics that could be considered to be covered by the general guidance on data analysis for ML development.

3. The EWG Chair indicated that in addition to the topics mentioned in document CX/CF 19/13/16 the following topics could be included for further consideration by CCCF.

   a) Importance that food and feed for which data are provided are correctly identified and reported with detailed information on the food or feed concerned (correct identification, state of the food/feed (fresh, dried, ready-to-eat, etc.)

   b) Handing of the data not provided to the GEMS/food

   c) Handling of outliers

   d) Handling of data for which it can be reasonably assumed that the unit of the data provided or the basis of the data are expressed (e.g. fat basis vs whole weight) is not correct.

   e) Lack of information on data provided

   Following the discussion at the 13th session, it was agreed that

   a) The scope of the work should be extended to guidance for improved data collection related to comments on the importance of detailed information to be provided with occurrence data and to reflect this extension of the scope in title of the document.

   b) To delete the criterion of the evaluation if provided occurrence data reflect the application of Codex Code of Practice and/or GAP/GMP as this was considered not to be feasible.

[1] REP18/CF, paras 155-156

CCCF noted the following topics for further consideration for guidance

- information on the methods of analysis and their validation used for generating occurrence data;
- handing datasets with a different contamination pattern (e.g. as consequence of originating from different regions, different production years);
- providing guidance on when to combine or keep separate such datasets for assessment;
- re-iteration of importance of providing sufficient detail of provided data to allow correct grouping. This correct grouping is also of major importance for correct use of these data for exposure assessment;
- include guidance on how to present data in EWG reports to CCCF.

5. A discussion paper for consideration by the established EWG has not been prepared. Therefore, this discussion paper contains the list of items, as discussed at the 13th session, prepared by the chair of the EWG, that could be the basis for a preliminary discussion at the current session, also taking into account the experiences with the analysis of the data for the establishment of maximum levels of cadmium in chocolates (agenda item 6), maximum levels for lead in certain food categories (agenda item 8) and maximum levels for total aflatoxins in certain cereals and cereal-based products including foods for infants and young children (agenda item 10 (a)) and this in view of a more elaborate document prepared by the EWG in view of a discussion at CCCF15.

6. Given the urgency and the importance of having available an elaborated document for discussion at CCCF15, the work in the re-established EWG will start very shortly after CCCF14.

RECOMMENDATIONS

7. The CCCF is invited to

a) Have a consideration on the appropriateness of the identified topics in Annex for inclusion in a guidance for data analysis for ML development and improved data collection, and in particular on the suggestion to include guidance on elements to consider to define the rejection rate to be applied in function of the product type and contaminants.

b) Have a non-binding consideration of other topics which would be appropriate to be included in a guidance for data analysis for ML development and improved data collection.

c) Agree to re-establish the EWG to elaborate the draft of a general guidance on data analysis for ML development and improved data collection taking into account the outcome of the discussion at this meeting for discussion at CCCF15 (2022).

d) Given the importance of this work for future discussions on MLs within CCCF, to urge the Chair of the EWG to start the work within the EWG without any delay and to report every two months, the first time on 15 July 2021, to the Codex secretariat on the progress achieved to ensure a timely completion of the guidance for discussion at CCCF15.

ANNEX

A) CRITERIA FOR THE ESTABLISHMENT OF MAXIMUM LEVELS IN FOOD AND FEED[2]

Selection of criteria has been made of relevance for improved data collection and analysis of data for setting MLs

- Validated qualitative and quantitative analytical data on representative samples should be supplied. Information on the analytical and sampling methods used and on the validation of the results is desirable. A statement on the representativeness of the samples for the contamination of the product in general (e.g. on a national basis) should be added. The portion of the commodity that was analyzed and to which the contaminant content is related should be clearly stated and preferably should be equivalent to the definition of the commodity for this purpose or to existing related contaminant regulation.
- Information on appropriate sampling procedures should be supplied. Special attention to this aspect is necessary in the case of contaminants that may not be homogeneously distributed in the product (e.g. mycotoxins in some commodities).
- MLs should be set as low as reasonably achievable and at levels necessary to protect the consumer. Providing it is acceptable from the toxicological point of view, MLs should be set at a level which is (slightly) higher than the normal range of variation in levels in food and feed that are produced with current adequate technological methods, in order to avoid undue disruptions of food and feed production and trade. Where possible, MLs should be based on GMP and/or GAP considerations in which the health concerns have been incorporated as a guiding principle to achieve contaminant levels as low as reasonably achievable and necessary to protect the consumer. Foods that are evidently contaminated by local situations or processing conditions that can be avoided by reasonably achievable means shall be excluded in this evaluation, unless a higher ML can be shown to be acceptable from a public health point of view and significant economic aspects are at stake.
- Proposals for MLs in products should be based on data from various countries and sources, encompassing the main production areas/processes of those products, as far as they are engaged in international trade. When there is evidence that contamination patterns are sufficiently understood and will be comparable on a global scale, more limited data may be enough.
- MLs may be set for product groups when sufficient information is available about the contamination pattern for the whole group, or when there are other arguments that extrapolation is appropriate.
- Numerical values for MLs should preferably be regular figures in a geometric scale (0.01, 0.02, 0.05, 0.1, 0.2, 0.5, 1, 2, 5 etc.), unless this may pose problems in the acceptability of the MLs.

[2] Reference is made to the criteria for the establishment of maximum levels in food and feed as provided for in Annex I of CXS 193-1995 General Standard for Contaminants and Toxins in Food and Feed

B) IMPROVED DATA COLLECTION

Important elements to be provided when reporting occurrence data

1) information on the stage in production and production chain where the sampling took place (farm, wholesale, import, retail) and location (country/region) of sampling. If known, origin of product sampled.

2) information on type of sampling : targeted sampling, suspect sampling, random sampling

3) food and feed to be correctly identified and reported with detailed information on the food or feed concerned (correct identification, state of the food/feed (fresh, dried, ready-to-eat, etc.)

4) information on the portion of food analysed (e.g. peeled or not, edible part or whole fruit, etc..)

5) the unit of the data provided and the basis on which the data are expressed (e.g. fat basis vs whole weight)

6) information on the methods of analysis (and their validation) used for generating occurrence data with information on the LOQ/LOD of the method

C) HANDLING/ACCEPTANCE OF DATA WITHIN A DATASET

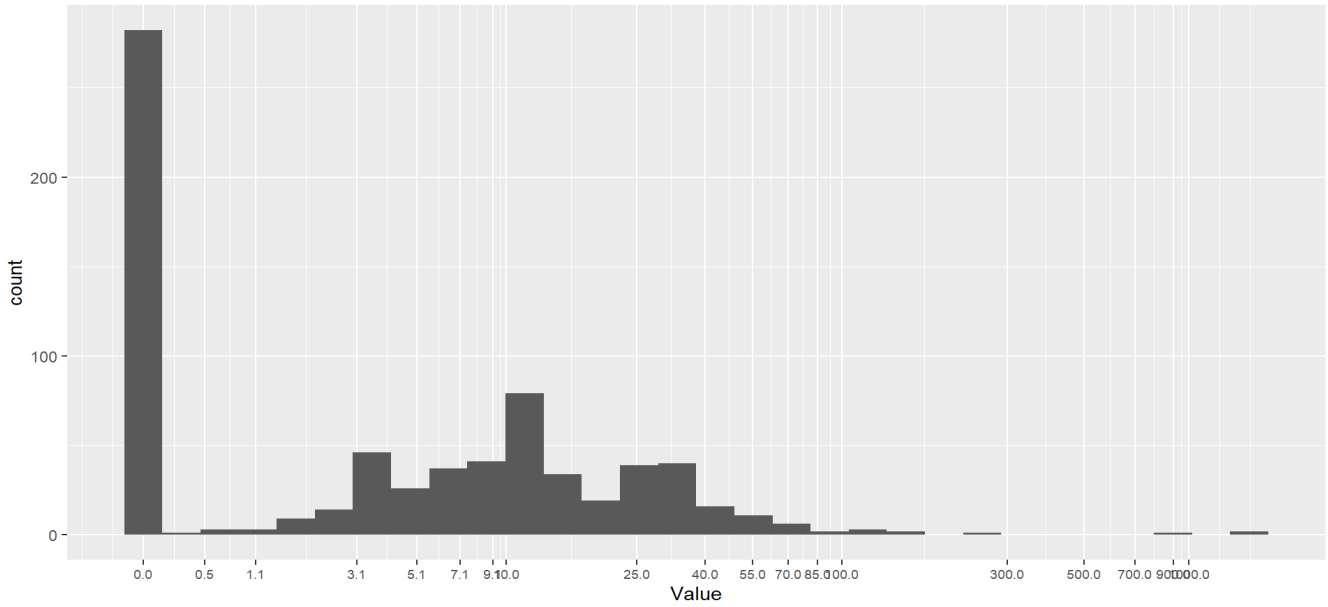
1) Handling of outliers/extreme values

When to consider data as outliers/extreme values?

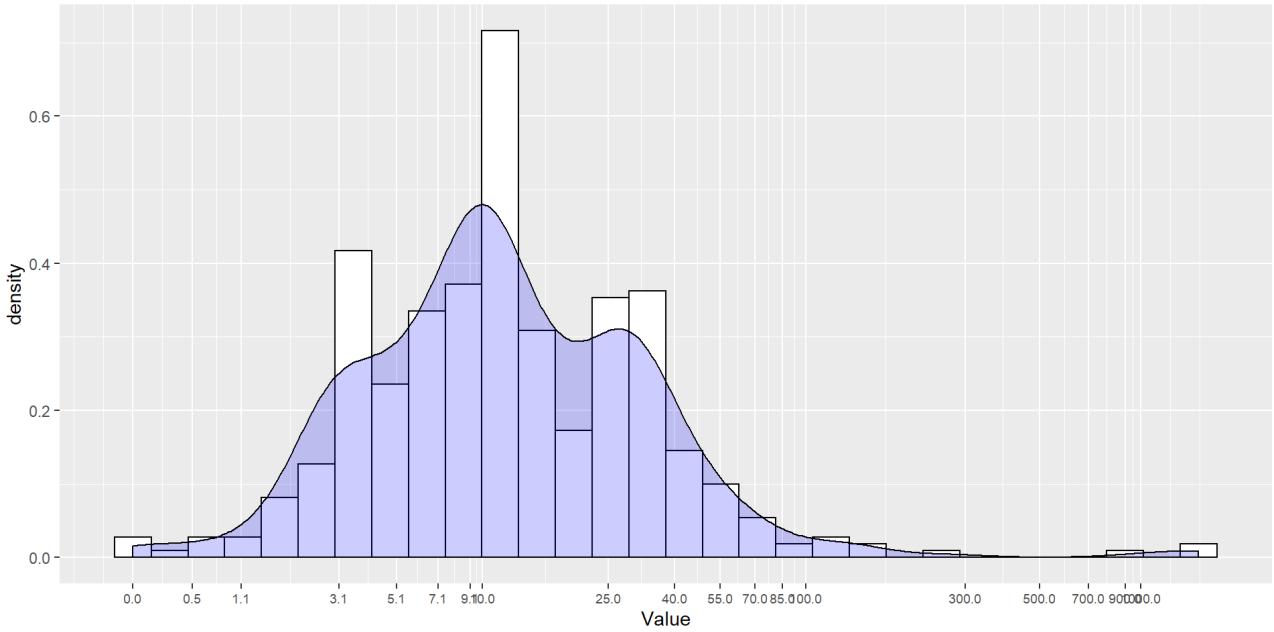
For consideration: In case the data are outside the range of distribution of the data and no justification can be provided for these extreme results (such as data from a year with extreme weather conditions, data from a specific region/continent, …)

Example EU data on sum of T-2 and HT-2 toxin in oat milling products (717 results of which 438 quantified results)

Histogram: all results



Histogram/density of quantified results



In case no justification for the data with levels > 500 µg/kg can be provided these data could be considered as outliers.

2) Handling of data for which it can be reasonably assumed that the unit of the data provided or the basis on which the data are reported (e.g. fat basis vs whole weight) is not correct.

If there are clear indications that the unit in which the data are expressed is incorrect or the basis on which the data are expressed is incorrect, these data should be excluded from further data analysis.

Examples of "clear indications"

* Levels within a data set of 200 results are in the range of 0 to 20. All data are expressed as µg/kg, except 5 quantified data points expressed as mg/kg. When putting these data in a frequency distribution curve (see a) they would be identified as possible outlier

* Levels from a food with a typical fat content of 5 % within a data set of 200 results of which all data are expressed on whole weight. 195 results are falling in the range of 0-20 mg/kg, however 5 data points are falling within in the range of 100 – 400 mg/kg. When putting these data in a frequency distribution curve (see a) they would be identified as possible outlier

3) Lack of information on data provided

It has to be considered to which extend the missing information makes the data unusable.

Examples of missing information by which data cannot be used for further data analysis:

- All data from a dataset are reported as < LOQ and the LOQ is not provided (more information in point 2 in Chapter D)
- the unit in which the result is reported is missing or the basis on which the result is expressed
- the state of the food sampled (dried fresh)

Examples of missing information but the data could still be used for further data analysis:

- sampling information: type of sampling, year of sampling, location of sampling, …
- method of analysis used

4) Handing of the data not provided to the GEMS/food

It has to be considered if these data can be taken into further data analysis

– in case there are only limited data available in the GEMS/food database, it could be considered useful to use these data in further data analysis.

– in case there are extensive data available in the GEMS/food database, it could be considered not to use these data in further data analysis (and certainly not in case the data do not show a contamination pattern different than the data available in the GEMS/food database).

5) Handling of datasets with a different contamination pattern (e.g. as consequence of originating from different regions, different production years)

Guidance should be provided on when to combine or keep separate such datasets for assessment

– if datasets from different regions/continent in the world show a different contamination pattern and a valid reasoning for the difference can be provided (e.g. different climate conditions, different production conditions/techniques) , then the datasets could be kept separate for assessment.

D) IMPORTANT TOPICS TO BE CONSIDERED FOR DATA ANALYSIS

1) Minimum number of samples needed for the use of percentiles

Background information

In order to apply the above criterion "*MLs should be set at a level which is (slightly) higher than the normal range of variation in levels in food and feed"*, high percentiles are used to define that level. The reliability of high percentiles is related to the number of data used to calculate them. Percentiles calculated on a number of subjects should be treated with caution as the results may not be statistically robust.

A clear indication concerning the minimum number of observations necessary to estimate a given percentile is not provided in literature. Different options can be used, none of them being a widely accepted standard.

A very simple option is to require that the calculated percentile must at least be different from the maximum value within the sample. This means that at least 20 observations are needed to identify the single observation at the 95^{th} percentile and 100 observations are needed for the 99^{th} percentile.

In statistics, the coverage probability of a confidence interval is the probability that the interval contains the true value of interest (e.g. 95th or 99th percentiles). When the number of observations is not large enough, the coverage probability may not attain the nominal value, and drops below, for example, 95%. This is more likely to occur at high percentiles, e.g. 95th or 99th. Therefore, the coverage probability has been used to set guidelines to determine the minimum number of samples for which (extreme) percentiles can be computed. In the case of significance level (α) being set at 0.05 to determine a 95% confidence interval, the coverage probability should target 95%. In this case, this is achieved for $n \geq 59$ and $n \geq 298$ for the 95th or 99th percentiles, respectively.

2) Limit of Quantification (LOQ) considerations

Several situations applicable to datasets provided can occur and the guidelines to be elaborated should provide guidance on how to handle the datasets in the different situations

- No LOQ provided
    - Dataset contains (nearly) all quantified results
    - Dataset contains a significant part of left-censored data (i.e. < LOQ) and no LOQ provided

      In the above situations where the LOQ is not provided, should the guidance provide for different conclusions as regards how to handle the dataset in case the quantified results (significantly lower than the ML under consideration) in the dataset provide a an indication that the LOQ is (very) low compared to datasets where the quantified results do not provide that indication.
- LOQ provided
    - Dataset with LOQ significantly lower than the ML under consideration
    - Dataset with LOQ in the range of the ML under consideration
    - Dataset with LOQ above the ML under consideration

In the above situations where the LOQ is provided, should there be guidance on cut-offs to be used for the LOQ on the analytical results dataset used for the ML development?

Should the guidance provide for different conclusions as regards how to handle the dataset in case the dataset contains nearly all quantified results compared to a dataset with nearly all left-censored data?

3) Using data sets with a large proportion of left-censored data for ML development

In certain cases, the analytical results for one specific contaminant are produced with a battery of different analytical methods and/or the same analytical method but with very different sensitivities. As a consequence, there could be a wide range of limits of detection (LOD) and limits of quantification (LOQ) for a particular contaminant and food matrix in a given dataset, composed of datasets from different sources. This situation is particularly relevant when the occurrence datasets used for the ML development contain a high number of non-quantified/non-detected data (left-censored data).

The standard approach to deal with left-censored data is the use of the substitution. In this method, at the lower-bound (LB), results below the LOQ and LOD are replaced by zero; at the upper-bound (UB) the results below the LOD are replaced by the numerical value of the LOD and those below the LOQ are replaced by the value reported as LOQ. Additionally, as a point estimate between the two extremes, the middle-bound (MB) scenario is calculated by assigning a value of LOD/2 or LOQ/2 to the left-censored data.

4) Geographical coverage of the provided occurrence data

Guidance should be provided to evaluate the appropriateness of the geographical coverage of the provided data for ML development and a procedure should be developed for situations for which it is concluded that the available data do not provide a sufficient/appropriate geographical coverage.

5) Period coverage of the provided occurrence data

Guidance should be provided in which situation it might be required that the provided occurrence data relate to several production years for ML development (can be different for different types of contaminants: mycotoxins, plant toxins, processing contaminants, environmental contaminants in function of the assumed year to-year variation or evolution of contamination in time)

6) Data sets with low number of data (e.g. less than 60) for development of ML

Guidance could be given in which situations it can be concluded that the data, despite the low number, are sufficient for the development of an ML (e.g. despite limited number good geographical coverage, no large variation in occurrence observed despite data originating from different regions/from different years, etc).

E) GUIDANCE ON HOW TO PRESENT THE DATA IN EWG REPORTS TO CCCF

It is important that the data are represented in such a way in the EWG report to CCCF to enable an informed discussion on appropriate MLs to be established.

The detail of reporting depends on the amount of data available and also of the nature of the contaminant.

Elements of consideration (not-exhaustive)

- if there is a significant year-to-year variation in occurrence it is appropriate to provide an analysis of the data per year.
- if there is a significant difference in contamination pattern between regions of e.g. climate conditions or production methods, it is appropriate to provide an analysis of the data per year;
- the description of the data should provide a clear view on the data set e.g;

* Number and proportion of positive (quantified results)

* Mean, median and range of positive results

* P90, P95, P99
* histograms/density of positive results (example see

F) ISSUES IDENTIFIED IN THE DATA ANALYSIS FOR POSSIBLE MLs OF LEAD (agenda item 8, CX/CF 21/14/8) and MLs of TOTAL AFLATOXINS (agenda item 10 (a) CX/CF 21/14/10 – Part I) NOT MENTIONED BEFORE

1) Application of different rejection rates for different types of products and contaminants, deviating from the usual rejection rate of 5%

At the 13th session of CCF it was clarified that the basis on which the MLs should be proposed (i.e. rejection rate, occurrence data and reduction risk) was outside the scope of the guidance (§ 162, REP19/CF)

However there is the explicit request to the CCCF in relation with the discussion on MLs for lead and total aflatoxins whether different rejection rates should be applied for different types of products and contaminants. Therefore, CCCF might agree that it is appropriate to provide in this guidance, elements which should be taken into account to define the appropriate rejection rate. This should increase the transparency on the basis on which grounds a maximum level has been set.

Possible elements for consideration (not exhaustive)

- nature of the product:
    - raw cereals of which already large part is used for feed: non-compliance with the food ML might not necessarily result in economic damage as it can still be used as feed.
    - processed products intended for human consumption: non-compliance with the food ML will result in economic damage as possible alternative uses will result in lower return or in certain cases the lot has to be destroyed.
- different regional contamination patterns:
    - worldwide dataset might have a rejection rate lower than 5 % at a certain ML while regional datasets might have for the same ML much different (lower or higher) rejection rate.