Workshop report

Preparatory work on how to report, use and interpret historical control data in (eco)toxicity studies

Tamara Coja[1,a], Agathi Charistou[2], Aikaterini Kyriakopoulou[2], Kyriaki Machera[2], Ulrike Mayerhofer[1], Dimitra Nikolopoulou[2], Eliana Spilioti[2], Anastasia Spyropoulou[2], Vasilis Souvlidis[2], Johann Steinwider[1], Tanja Tripolt[1]

[1] AGES, Austrian Agency for Health and Food Safety, Austria

([a] coordinator, corresponding author)

[2] BPI, Benaki Phytopathological Institute, Greece

Abstract

EFSA has granted (GP/EFSA/ENCO/2020/02) a project called "Preparatory work on how to report, use and interpret historical control data in (eco)toxicity studies" with the aim to collate all relevant information from public literature, as well as stakeholders experience, knowledge and understanding of the use and interpretation of historical control data (HCD) when evaluating toxicity studies. One of the activities in the project concerned the organisation of an international workshop, involving all relevant stakeholders. The workshop has been organised by the beneficiaries of the project, Austrian Agency for Health and Food Safety (AGES) and Benaki Phytopathological Institute (BPI) from May 3-5 2022, as a virtual event. The workshop was addressed to all stakeholders with expertise in the area of conduction and interpretation of toxicity studies on mammals and/or expertise in human risk assessment. The use of HCD in context of ecotoxicity studies (other than conducted on mammals) was subject to limited discussions, but it was agreed that general principles are the same as for toxicity studies; the only difference is in the potential interpretation of results. The intense discussion during the workshop revealed the need for a clear set of criteria to be fulfilled before HCD can be taken into account, for high level of granularity in the presentation of HCD to understand potential differences between control populations and for close exchange between different domains (statistics, toxicology) before appropriate analysis is applied. In order to harmonise the use, reporting and interpretation of HCD within and between national regulatory bodies and European Agencies (EFSA, ECHA), the establishment of a set of relevant criteria/guidance was recommended.

Key words: historical control data, genetic drift, statistics, histopathology, reproductive toxicity, clinical pathology, harmonisation, guidance

Disclaimer: The summaries of presentations and panel discussions reflect the outcome of the workshop recordings and no personal opinion of the authors of the workshop report.

Acknowledgements: Workshop organisers wants to thank cordially to all speakers and chairman persons, which supported the workshop during the preparation and implementation phase (in alphabetical order): Thomas Bean (FMC), Manon Beekhuijzen (Charles River Laboratories), Diane Benford (Vice chair of EFSA Scientific Committee), Chad Blystone (National Toxicology Programme), Adeline Cavelier (ANSES French Agency for Food, Environmental and Occupational Health & Safety), Peter Clausing (PAN Germany), Jutta Davidson (Charles River Laboratories), Stephen Dertinger (Litronlabs), Bertrand Desprez (French Agency for Food, Environmental and Occupational Health & Safety), Manousos Foudoluakis (Corteva), Sibylle Grötters (BASF, on behalf of RITA), Thomas Hofmann (BASF, on behalf of Crop Life Europe), Urte Jäh (Charles River Laboratories), Laura Martino (European Food Safety Authority) Aidan McGuire (Charles River Laboratories), Chiara Perazzolo (European Chemicals Agency), Susanne Rudzok (German Federal Institute for Risk Assessment), Volker Strauss (on behalf of ESTP). Manuela Tiramani (European Food Safety Authority), Martin Wilks (University of Basel).

Suggested citation: Coja T, Charistou A, Kyriakopoulou A, Machera K, Mayerhofer U, Nikolopoulou D, Spilioli E, Spyropoulou A, Steinwider J, Souvlidis V, Tripolt T, 2022. Workshop report: How to report, use and interpret historical control data in (eco)toxicity studies.

Table of contents

1. Introduction

This report summarises the outcome of a scientific workshop on How to report, use and interpret historical control data in (eco)toxicity studies, held as virtual event from 3rd to 5th May, 2022, 12:00-16:00 Central European Time (CET).

EFSA has granted (GP/EFSA/ENCO/2020/02) a project called "Preparatory work on how to report, use and interpret historical control data in (eco)toxicity studies" with the aim to collate all relevant information from public literature, as well as stakeholders experience, knowledge and understanding of the use and interpretation of historical control data (HCD) when evaluating toxicity studies.

One of the activities in the project was the organisation of an international workshop, involving all relevant stakeholders. The workshop has been organised by the beneficiaries of the project, Austrian Agency for Health and Food Safety (AGES) and Benaki Phytopathological Institute (BPI).

The workshop was organised as an information gathering event, with the aim to explore the landscape on use, reporting and interpretation of HCD. The workshop offered the platform for discussion on different aspects, like biological variations, environmental parameters, pathological terminology, statistics, *etc*., all relevant for better understanding how to use HCD in the interpretation of the effects observed in toxicity studies. It reflected the needs of scientists working in the area of toxicological risk assessment, as captured in the survey, launched from January 17 – February 21, 2022, specifically conducted as one additional activity within the project.

The workshop was addressed to all stakeholders with expertise in the area of conduction and interpretation of toxicity studies on mammals and/or expertise in human risk assessment. The use of HCD in context of ecotoxicity studies (other than conducted on mammals) were subject to a very limited time slot.

Hundred eighty two (182) registered participants from three continents (Europe (24 countries), North America (Canada, United States), Asia (China, India, Lebanon, Japan)) (Annex D) attended the workshop. The participants were mostly from regulatory authorities (57%) and from industry (19%), followed by contract research organisations (13%) and consultants (7%).

The workshop programme was structured in three main sessions, on three consecutive days, covering the key topics related to HCD (Annex A). Each main session consisted of presentations (Annex B) in which the themes were introduced, followed by a panel discussion with chair facilitation where participants could ask questions to the speakers.

This report follows the general structure of the workshop. Relevant outcome of the discussions, as well as topics for further consideration are summarised in the conclusions and recommendations.

Following the workshop, registered participants were asked for their feedback (Annex C). Forty four (44) answers were received to the questions a) if the workshop covered all relevant topics in relation to HCD and b) which are the major issues identified by participants for further consideration. Ninety-one percent (91%) of participants stated that all topics relevant in context of HCD were covered in the Workshop.

Based on the survey and the workshop as preparatory tasks, the EFSA Panel on Plant Protection Products and their Residues (PPR) will draft a Scientific Opinion (https://open.efsa.europa.eu/questions/EFSA-Q-2021-00274).

2. Workshop

2.1. Summaries of presentations

2.1.1. Day 1 of the workshop

Standardisation and Biosecurity in Laboratory Rodent Breeding (Jutta Davidson, Urte Jäh, Charles River Laboratories)

The presenter pointed out that standardization in laboratory rodent breeding requires four key elements in order to ensure reliable data and reproducible scientific results in research: Genetics, Health, Environment and Handling. All these key elements are important factors with regard to standardization, however, due to the complexity of this subject this presentation focused mainly on genetics.

It was noted that genetic stability programs should ensure genetic stability over generations and breeding locations (genetic management system), while (unconscious) selection, mutation and genetic drift are factors that have an impact on the genetic identity. An approach, employed at Charles River for outbred strains, to maintain the genetic diversity is the International Genetic Standardization (IGS). IGS is a genetic management system for organizations to be able to exchange control data with the same genetic standard worldwide. IGS uses pedigreed gnotobiotic foundation colonies, a large population size, equal distribution of genders, avoidance of inbreeding in production colonies prevented by rotational breeding programs and a program of regular breed stock migration between sub-populations (to and from foundation colony which is the genetic standard of a certain outbred stock model).

Apart from genetics, important microbiological, environmental and handling standardization options were summarized.

It was highlighted that the standardization actions taken by breeders concern only approx. 5% of the lifespan of rodents, i.e. until shipping.

National Toxicology Program's Perspective and Use of Historical Control Data (Chad Blystone, National Toxicology Program)

National Toxicology Program (NTP) considers the concurrent control as most important for the evaluation of study findings. HCD are regarded as an important resource for evaluation of current datasets, but it is acknowledged that they are influenced by various factors like genetics, route of exposure, nomenclature and study design. Within NTP, standardization of nomenclature and diagnosis is achieved *via* a specific review process for carcinogenicity studies, where any inconsistencies are resolved *via* a Pathology Working Group.

While NTP provides HCD for neoplastic lesions and development and reproductive toxicology (DART) studies, no HCD are provided for non-neoplastic lesions. However, NTP provides an atlas for non-neoplastic lesions which contains guidelines for microscopic non-neoplastic lesions in rats and mice.

Currently, HCD from carcinogenicity studies and DART studies are collated and published on the NTP website in a rolling 5-year window, based on the date of the study start (incidence rates of neoplasms, growth, survival data provided by species, strains and sex, route of exposure and diet).

Reporting of neoplastic lesions (by organ system) includes individual findings, total, mean and standard deviation. In some cases, also combinations of neoplasms are provided, e.g. adenomas and/or carcinomas.

Relevant changes and challenges in NTP HCD (change in rat strain (Fisher to Sprague Dawley), in exposure type (exposure starting in utero for carcinogenicity studies), in histopathology (number of brain sections investigated)) were highlighted. HCD are an invaluable resource for evaluation of current datasets requiring, however, consistency in methodology, nomenclature, and use. It was also stressed that the progression in methods to increase sensitivity and the standardization to increase consistency lead in fracturing of historical control databases, and thus, changes need to be planned ahead of time.

Use of HCD on Pharmaceutical Toxicology Studies at Charles River Edinburgh (Aidan McGuire, Charles River Laboratories (CRL))

There is no requirement to add HCD ranges to pharmaceutical toxicology reports (differently than for agrochemicals) and it was stated that at CRL HCD are used rarely in preclinical tests in pharmaceutical toxicology, where data from the contemporaneous control group is given precedence. HCD are used only to explain high individual values, usually noted in the control animals, that cause artificial statistical significance. Thus, with the exception of specific carcinogenicity (tumor incidence) data, HCD may be used to confirm that an observed change is by chance and not treatment related.

Several provisions for standardization of clinical pathology data collection were mentioned, e.g. same site and analytical methodology, known fasting status, recording of anesthesia, data must be contemporaneous.

Reporting of clinical pathology data at CRL comprises the supplier (if requested), strain, age range, parameter quoted, absolute numbers of animals, mean, median, minimum, maximum, 95% confidence limits and standard deviation.

It was stressed that HCD data should be used from the same site and the studies should be conducted under the same conditions. While the supplier makes only little difference for clinical pathology parameters in dogs, housing has a significant impact on clinical parameters, e.g. reduced stress by group housing, which improves the quality of the study. Fasting rats overnight was shown to cause expected changes in glucose and triglyceride levels and minor changes to liver enzymes and electrolytes. However, there is no evidence that the variability of data from fasted animals is less compared to non-fasted animals. Clinical chemistry data in Han Wistar rats of Charles River have shown that Isoflurane anesthesia has a bigger impact on glucose levels than fasting.

Challenges in Using the HCD: a methodological perspective (Laura Martino, EFSA)

Two methodological approaches for the use of HCD were highlighted in line with the main results of the survey: the establishment of Reference Intervals (RI) from HCD and the integration of HCD in statistical analyses to draw conclusions.

Reference intervals (RI) were explained to be a possible tool for the assessment of the reliability of the concurrent control as well as to identify potential outliers. However, the comparability of the control individuals to the HCD and the availability of HCD at individual level/quantiles of distribution were considered as critical issues with regard to RI. Variability (including intra/inter-individual variability, variability across studies) and bias were highlighted as two relevant challenges. Possible strategies to reduce variability and bias were discussed, which comprised subgrouping of the RI to factors that are sources of variability, the use of weights (0-1) reflecting similarity of the individuals from HCD and the concurrent controls, performance of a sensitivity analysis and careful removal of the individuals potentially affected from bias. Furthermore, the necessity to get sufficient precision for the estimation of extreme centiles to reduce sampling uncertainty to an acceptable level was stressed.

Reference was made to the outcome of the survey conducted within the project. More specifically, it was noted that in the survey the responders consider as promising the use of HCD to establish RI or integrate the HCD in statistical analysis to draw conclusions. The use of the RI established from relevant HCD in assessing the reliability of concurrent control or identifying potential outliers was further highlighted. It was recommended that the equivalence limits should not be too wide, because this could lead to wrongly concluding towards equivalence when it is not true (i.e. false positives).

As another option it was suggested to use HCD to establish priors on the parameters of interest when adopting a Bayesian approach. Additionally, it was suggested to give consideration to the use of weighing factors to reduce the influence of HCD. The application of a sensitivity analysis is recommended.

It was recommended that ideally, HCD should be provided at individual level including metadata. Alternatively, the following minimum requirements were suggested: mean, standard deviation, minimum and maximum of the HCD distribution both overall and by subgroups, median, quartiles and extreme centiles (e.g. 2.5% and 97.5%) of the HCD to assess the distribution of the HCD and the comparability with concurrent controls and to identify outliers in the concurrent controls. For graphical representation of HCD, boxplots and histograms were suggested.

Three main messages were given: 1. to be useful, HCD needs to meet minimal requirements in terms of format, accuracy, sample size required to derive RI with a pre-defined precision; 2. Providing adequate description of metadata describing how HCD have been collected is key for a proper use; 3. It is important to plan upfront in a protocol the expected use of the HCD to avoid bias in the conclusions; 4. Open access to HCD would provide great benefit to the scientific community.

According to the results of the conducted survey before the workshop it was concluded that a guidance on the use, report and interpretation of HCD as well as a harmonized terminology and a large set of studies to source HCD are needed.

2.1.2. Day 2 of the workshop

Historical Control Data in Pathology: Meaningful Use and Limitations (Sibylle Gröters, on behalf of RITA initiative)

Standardization of diagnostic criteria and data generation was explained to be pivotal, which can be ensured by inhouse standard operating procedures (SOPs), international harmonized guidances, standardization and comparability of slides, standardized terms, INHAND criteria (International Harmonization of Nomenclature and Diagnostic criteria) and the goRENI system (global open Registry Nomenclature Information System).

The Registry of Industrial Toxicology Animal-data (RITA) Project, founded in 1988, was introduced. Currently, eight companies from Europe and USA participate in this project with the purpose to generate harmonized data by centralization of consistent HCD from rodent carcinogenicity studies from different laboratories. Cross organizational reviews of studies contribute to high reliability, robustness and quality of histopathological assessments of tumors.

It was pointed out that pathological evaluation needs training and education of pathologists both in diagnostic as well as toxicological pathology. As regards the Pathology Peer Review, different steps within this process were explained. Basically, the study pathologist examines all slides and drafts a pathology report. If necessary, there might be an internal and/or external Pathology Peer Review of selected slides (second/third opinion). In case of any questionable findings, selected slides may also be

examined by a pathology working group leading to an expert opinion which is binding and accompanies the final report.

Types of HCD in pathology comprise descriptive findings (incidences in histopathologic findings non-neoplastic, pre-neoplastic and neoplastic lesions) as well as findings like organ weights, differential ovarian follicle count, morphometry, cell proliferation, etc.

Various aspects with regard to HCD were discussed, i.e. for non-neoplastic lesions the lesion might be diagnosed differently in various study types (e.g. 28 day study vs. carcinogenicity study). Furthermore, the 5 year time frame for compilation of HCD was considered critical since the number of studies in the required period is rather low for long-term and carcinogenicity studies. In addition, it was stressed that for rare tumors more studies are necessary to be included in the HCD to better understand the distribution and cover greater control population. In addition, challenges in the evaluation of effects on organ weight, including high variation in the weight of reproductive organs in dogs (different start in puberty, inter-individual differences and long estrus cycle) were discussed.

It was noted, that not all organs and all species could be judged the same way. First choice of control data is the concurrent control, followed by in-house data, RITA database and other sources (NTP, literature, CRO data). Minimum, maximum, mean and standard deviation (only for true numbers) as well as the performance of a trend analysis in some cases (tumors over the years) were considered as necessary reporting endpoints.

How to report, use and interpret historical control data in DART studies (Manon Beekhuijzeen, Charles River Laboratories)

It was pointed out that HCD should be considered as a tool for a better understanding of events or apparent differences observed within a study, while it should not be used for discounting unwanted or difficult findings. HCD were regarded to be more relevant in screening/pilot studies, since only a small number of data from concurrent controls is available.

It was stated that HCD can be used in three primary ways: identification of a) aberrant control values as well as understanding the relevance of b) low- and c) high-incidence findings.

Regarding the requirements for HCD, it was highlighted that HCD should contain studies conducted with the animals of the same strain, age, sex, from the same animal supplier and the laboratory as well as collected within a 5-year period. In addition, for DART studies the international harmonized terminology developed by the IFTS (International Federation of Teratology Societies) should be applied. Continuous data should be reported as mean for each litter. For binary data, the proportion of fetuses affected in the litter should be calculated.

It was noted that the litter should be used as the experimental unit in DART studies, because same finding in fetuses from several litters are of greater concern than same findings in several fetuses of the single litter. Furthermore, using the fetus instead of the litter could lead to invalid statistics due to inflation of the group size. It was emphasized that statistical analysis alone cannot be used to decide whether the study finding is a treatment related effect, because the study may not have the necessary statistical power or because of the possibility that observations per study can attain statistical significance by chance alone. Therefore, the biological significance of statistically significant findings should be evaluated based on e.g. dose-dependency and HCD. In addition, the importance of using the same vehicle and volume when considering HCD was stressed. It was pointed out that e.g. hormonal fluctuations (e.g. high variability of T4) have to be taken into account, since there is no pretest value and only one measurement is taken within a toxicity study. Consequently, in these cases HCD should be considered in order to decide if there is a possible treatment related effect.

Establishment and Use of Historical Control Data in Clinical Pathology (Volker Strauss, on behalf of ESTP)

The main issues regarding the confidence in HCD for clinical pathology were described to be differences in study conduction. The establishment of historical control intervals according to regulatory guidance, proceeding according to documented procedures, the application of reasonable statistical methods and an objective interpretation of study values were stressed to be taken into account.

Other than specified in some guidance (e.g. JMPR 2015) which recommend a time interval for HCD of 2.5 years before and 2.5 years after the actual study, the industry standard is to use HCD from 5 years before the actual study, based upon the practicality (when the HCD are to be included in the final study report, no additional 2.5 years period can be awaited). The application of a greater time interval was proposed in case there are less than 10 studies or less than 20 individual values including a trend analysis that confirms that there is no trend among the HCD. Also merging of studies conducted via different routes of exposure was mentioned where it can be proven that route of exposure has no impact under the parameter under investigation.

The concurrent control group of the study or pre-study values are regarded as most important data for comparison. It was recommended not to use HCD established in external labs or textbook data for arguing that significant changes in toxicity studies are within the normal variation. As regards procedures of establishing HCD the possibility of collecting individual values and comparing mean and percentile range (confidence intervals) with study values or collecting means/medians of prior control groups (HCD interval established with minimum/maximum values after exclusion of outliers) were described.

Assessing the Quality of Historical Control Distributions and Calculating Useful Intervals: Genetic Toxicology Examples (Stephen Dertinger, Litronlabs)

Two examples related to *in vivo* micronucleus test and *in vivo* comet assay were discussed.

Generally, HCD can be used as component of demonstrating laboratory proficiency and study validity and it also can be used to judge whether study response data are clearly negative or clearly positive in addition to a pair-wise test that considers the concurrent control and a trend test. Negative HCD are used to show if the study results fall above or below an upper bound limit value of HCD. It was stated that judgement whether study response data are clearly positive or negative with HCD can be useful, in case it can be shown that the variability is primarily explained by biological variability. However, HCD are considered less useful when the variability is primarily caused due to inter-study nuisance factors (e.g. poorly controlled tissue harvest times, electrophoresis conditions).

It was pointed out that the quality assessment of the historical control distribution is necessary as a first step and more emphasis should be put on this, e.g. by using qualitative and semi-quantitative assessments, control charts, Stability Index and Variance Component Estimates. Qualitative and semi-quantitative assessments may serve to evaluate for example whether the data are consistent with published results from labs, whether the level of variation of samples within and across studies is comparable and whether there is an obvious drift. Using the Stability Index, the process variability can be considered. Major sources of variation can be quantified by applying the Variance Component Estimates via Residual Maximum Likelihood (REML), Nested Anova and Bayesian models.

If an assay has been found to be under control, the determination of intervals that describe the distribution of HCD is recommended.

It was suggested to use HCD in a flexible and nuanced manner and the need for a more detailed and transparent reporting of HCD and the type(s) of quality assessments was pointed out to gain confidence in their use.

2.1.3. Day 3 of the workshop

ANSES experience - pesticide evaluation in the area of mammalian toxicology (Adeline Cavelier, Bertrand Desprez, ANSES)

It was pointed out that the concurrent control data should always be the primary data to be used for interpretation of study results. HCD were regarded as one line of evidence for assessment and interpretation of toxicological studies, especially for effects like histopathological findings (incl. tumors), malformations, developmental parameters, gestational parameters, functional reproductive parameters and genotoxicity studies.

As regards evaluation of toxicity studies, it was indicated that HCD are mainly used for the assessment of the reliability of studies as well as for the interpretation of rare findings, borderline findings and genotoxicity studies since HCD represent data from a larger group size compared to the concurrent control group and thus give a better overview of the biological variation. It was pointed out that while HCD should always be used in a Weight of Evidence approach, they should not be used to dismiss adverse effects.

The use of the same species, strain, sex, laboratory, route of administration, study type and duration as well as the sufficient number of studies were regarded as minimum requirements of HCD. In addition, it was highlighted that HCD should be centered +/- 5 years around the date of the index study and the reporting of HCD should comprise at least mean, minimum, maximum, standard deviation as well as litter and fetal incidences for developmental toxicity studies. If these requirements are not met, HCD are not considered relevant or of low relevance.

To improve the relevance of HCD the following information was suggested to be included in the submitted dossiers: breeder, age and weight of the animals, information on environmental factors in the study, diet, vehicle, diagnostic criteria, staining method (developmental toxicity studies), standardized terminology, OECD and GLP status.

As regards the reporting of HCD it was recommended to include detailed statistical evaluation as well as the individual data. It was pointed out that mean and the range (minimum, maximum) are not sufficient since this allows no information on the distribution of the data. Therefore, a graphical presentation of the distribution was suggested. It was noted that the calculation of confidence limits should be harmonized.

Examples concerning the following issues as well as the respective solutions with regard to HCD were presented: poorly summarized HCD, using test guideline recommended 95% confidence limits, variability considerations on HCD by using 95% confidence intervals on percentages or proportions, getting all individual data.

It was concluded that there is a need for harmonization of the use, the minimum requirement of data provision and the reporting of HCD.

Experience from the European pesticide, biocide and C&L evaluation – Competent Authority perspective (Susanne Rudzok, BfR)

Data requirements for biocides and plant protection products were summarized. It was pointed out that basically, HCD should be taken into account if available. However, according to footnote 6 of the BPR Regulation 2012, the relevance of HCD has to be questioned in case of absence of any comparability between the study conditions.

Several challenges related to HCD were identified by means of relevant examples.

Frequently, there are issues with regard to the availability and representativeness of HCD as well as fulfilling the expected criteria (HCD should be centered as closely as possible to the date of the index study within a 5-year period, sufficient number of studies).

In addition, data gaps were highlighted to be another relevant issue (e.g. strain, weight, age, information on infections, only summary data, no appropriate data to compare incidences of malformations etc.).

Furthermore, it was pointed out that the interpretation of HCD poses a challenge. For example, HCD are often used to question treatment-relation of experimental results in the treated groups and to lesser extent to evaluate the reliability of the concurrent control in the study.

Finally, it was described that the assessment of the relevance of HCD for the observed effect is a challenge. In the end, HCD are only one aspect in the Weight of Evidence evaluation of the observed effect.

Historical control data in CLH dossiers (Chiara Perazzolo, ECHA)

Guidance on the application of the CLP criteria, 2017 was summarized with regard to HCD, where it is stated that HCD should be used to check the validity of the concurrent control as well as to check the increase of rare tumours. HCD should be from the same strain/species and ideally from the same laboratory. It is also stated that HCD should be contemporary to the index study (e.g. period of 5 years). According to the CLP Guideline, HCD are also regarded as useful information with respect to the normal pattern and range of tumour types and incidences. However, it is stated that HCD should be on a case by case basis with consideration of their appropriateness and relevance.

Yet, it was indicated that these requirements are often not fully met in the submitted dossiers. The main problem that was highlighted is that often, the CLP dossiers only include ranges (minimum and maximum incidence).

Various positive and negative examples concerning the use and issues of HCD were presented (validity of the concurrent control, time range, HCD appropriateness and relevance, HCD from different laboratories).

It was concluded that the use of HCD in CLH dossiers did improve over time, but often only ranges of HCD are submitted which is not sufficient to assess the HCD. The assessment of the validity is often not explicitly stated and another issue that was pointed out is the use of HCD in addition to the concurrent control.

Acceptability and Use of Historical Control Data in Toxicological Studies (Thomas Hofmann, on behalf of Crop Life Europe)

HCD in regulatory toxicology were described to be used to assess biological variability through an increased power as compared to relatively small concurrent control groups, which is nevertheless regarded as the most important comparator. Furthermore it was stated that HCD help to reduce the number of control animals and thus contribute to the animal welfare. Moreover, it was summarized that HCD are used as quality assurance for the test system, for identification of aberrant control groups and distinguishing true response from incidental findings, to judge the relevance of findings and to address the statistical multiple comparison problem.

It was stressed that HCD should not be used with the intention to dismiss true findings.

The length of the time period covered by HCD was postulated to depend on the data type (continuous data or binary data). While the time period of 5-year (retrospective) was regarded as advisable for continuous data, also an adequate number of studies should be covered. Concerning binary data it was pointed out that a restriction to a 5-year period for HCD would lead to an information loss, since data on a distinct finding are not necessarily always observed (especially information on rare tumours or malformations).

In relation to the data type, HCD should also be dependent on the frequency of the finding and the number of studies available.

The reservations regarding the combination of studies with different routes of administration (feeding and gavage) were questioned but no data analysis is available yet.

Experience from the European pesticide evaluation – NGO perspective (Peter Clausing, PAN Germany)

It was pointed out that the concurrent control is regarded as most important data for comparison of study findings and that transparency is pivotal. It is required to use HCD from the same lab, same strain and same study duration within the last 5 years prior to the study.

Issues regarding HCD from NGO perspective were highlighted (e.g. information on start and end of the in-life phase is not stated in the renewal assessments and sometimes the date of the study report is missing, reporting of HCD, description of the HCD-source and time period of HCD to the index study).

It was stated that HCD should only be used in case the concurrent control data are questionable. In case of reassessments of old studies, HCD should be carefully used. Furthermore, the importance of details to ensure comparability was highlighted (e.g. influence of stressful housing on the immune system, genetic background).

It was suggested that reporting should include median and the interquartile range if there is a sufficient number of studies.

Avian reproductive toxicity studies – ecotoxicologist perspective (Manousos Foudoulakis (Corteva), Thomas Bean (FMC))

It was indicated that a key principle in ecotoxicology is to understand and define the baseline for a standardized test design.

Based on examples from publications it was shown that the main source of variability is due to variance within studies (intrinsic biological variation), while variance between studies and the source of birds

contribute significantly less. Case studies were presented, indicating that the determination of true treatment-related effects can be challenging. It was stressed that HCD may be a useful tool in the interpretation of ecotoxicological studies.

The comparison of data from 2005-2019 showed good agreement among labs for most endpoints. However, it was pointed out that data from the same laboratory is more useful than from other laboratories for birds. It was suggested that HCD could be pooled across labs based on the endpoint.

It was pointed out that there is a wide range of normal responses and the distributions may vary to a large extent, which requires careful interpretation. HCD can remove bias by providing context to the findings. However, the biological relevance and inherent measurement variations should always be considered.

Questions were raised regarding an appropriate time span for HCD and the number of studies that are needed. It was suggested that it would be best to have HCD from more than 20 studies, centered around the date of the index study. Also a reality check, using several time spans (e.g. +/- 2, +/-3 and +/-5 years) was recommended to ensure that there is no relevant influence by the size or the time span of the HCD.

It was concluded that for birds HCD should be more routinely used to aid data interpretation and decision making in a holistic assessment.

2.2 Summaries of panel discussions

During the workshop, there were few general agreements on use, report and interpretation of HCD:

- **Concurrent control group** is the most relevant control group for analysis of study results.
- The major requirement is that **comparability of relevant conditions** must be established/addressed before HCD are used for the analysis. From the statistical point of view there is the assumption that HCD and concurrent control come basically from the same population and are exposed to comparable environmental factors.
- When it comes to interpretation of study results, distinction between **relation to treatment** and **adversity** of a finding is crucial. **HCD can be applied as one line of evidence** to support interpretation of observations.
- **Specificity of an endpoint** has to be taken into account when evaluating HCD, as well as **justified statistical approaches** as chosen.

For different specific topics, very intense and fruitful discussion was recorded during the workshop. Discussed topics are summarised and most relevant outcome highlighted.

The genetic drift in the laboratory population can be minimised if the provider has a genetic stability programme in place. The genetic stability programme has started in the 1990-ies for rats and in 2001 for mice. If a genetic stability programme is in place, 5 years period for consideration of HCD is justified for outbred strains, while inbred strains might need special consideration. While outbred strain colonies under genetic stability programme are established with at least 100 or more non-sibling breeding pairs, inbred strains might need refreshing with frozen embryos, since less genetic stability is expected. Genetic drift is controlled and monitored over the time. It is also recognised, however, that same genetic material (as a starting point) is only one aspect of the standardisation, since only 5% of rodent life is covered by the time spent at breeding facility.
No tumour type has been highlighted to be more driven by genetic drift than the others.

The sample size needed to establish **reference ranges** strongly depends upon the endpoint, variability of the data and desired level of precision. As regards variability in the control group, where unexpected high values are noted, individual data can be investigated to eventually explore the cause.

When it comes to **statistical analysis**, the effect size should be established upfront the experiment but the whole study has to be taken into account when decision on adversity is made while evaluating the study. There are different methods to establish confidence intervals and the suitability of the method can be proven by sensitivity analysis (by repeating the analysis and changing one component or individual). For application of statistical methods an intense exchange between different domain experts (statisticians and toxicologists) is necessary.

Even if there is no general agreement if mean, full ranges or interquartile ranges, etc. are the most suitable statistical parameters, it is important to understand the distribution of data, also in the longitudinal way (over the years). This can be best achieved with larger datasets but it also depends from the variability in the response. In addition, the nature of the endpoint may influence the relevant statistical evaluation, e.g. for histopathological findings rather incidences are reported (1/50, 2/50, etc.) and ranges given, to see minimum and maximum number of affected animals per study.

As regards **outliers**, it was discussed how many studies/data points would be necessary to perform an outlier test and it was concluded that one size does not fit all. For clinical pathology, for example it was mentioned that 10 studies at least would be needed for the purpose of building a robust HCD to have confidence in identification of outliers. Different statistical methods might be appropriate for identification of outliers, also pending the parameter (e.g. parameters with zero counts like absolute basophil count might need different statistical method than parameters with high and variable values like liver enzymes).

While for histopathological findings exclusion of outliers is never done, in genetic toxicology the OECD TGs encourage the evaluator to exclude outliers (eliminate aberrant data) and trim the dataset. This is based upon the fact that if mutation happens in an early progenitor cell, unusually high frequencies are the consequence and make data interpretation difficult.

In any case, before a data point is considered as outlier, the reason for the value should be explored.

No globally agreed or standardised criteria for integration of HCD in the evaluation of the contemporaneous study (e.g. standardised "weighing" if HCD fulfil the requirements) exist. It is recognised that any analysis will strongly depend from the endpoint in question. While organ weight measurements may be better objectively compared when simply the age of animals is known, clinical pathology and haematology data, for example strongly depend upon the measuring equipment used, as well as from e.g. the time-point of blood collection or method of anaesthesia.

As regards **publicly available historical control data**, these are presented mostly as summaries on respective proprietary websites. It was mentioned by NTP and Charles Rivers Laboratories that individual study data from historical control databases can be requested for a certain time period, with precise information on what is needed. For every potential use of HCD there must be awareness of the procedures applied. As regards e.g. NTP data, change in number of brain sections from 3 to 7 could change the incidence in the data, as well as changes in study duration/exposure period in rat long-term studies.

There was discussion on **Open Access Global database**, and it was recognised that open access should not be mixed up with transparency of data. The temptation to misuse data of different sources, if publicly available, was discussed versus advantage of accessible and wide database. The comparability of HCD to concurrent control was highlighted as of utmost importance.

There was very detailed discussion on **histopathological evaluations for neoplastic findings**. It was emphasized that study results cannot be simply changed by the external pathologist (this is always a separate expert opinion) and incorporated in the final study report. If the laboratory internal peer review of histopathological findings is done (always conducted in carcinogenicity studies) this is described in the study report. For severity grading **INHAND** (International Harmonization of Nomenclature and Diagnostic Criteria for Lesions) gives criteria but the findings do not always follow the criteria as described. Such findings are then discussed in the panel meetings of the RITA initiative (Registry of Industrial Toxicology Animal-data).

RITA group of pathologists has up to 50 participants from 8 contributing companies, with 10-15 participants per panel meeting. When data are submitted to RITA, details on HCD like strain, breeder, test facility, age of animals, housing conditions, etc. have to be detailed. In case a report from RITA is desired, specific studies can be selected which fulfil the selection criteria and different level of granularity can be chosen.

If a finding is evaluated by the RITA panel meeting pathologists, the outcome does not change the final study report, it always stays an external expert opinion. RITA database contains mostly neoplastic findings, but also increasing number of findings in studies of shorter duration and non-neoplastic findings.

As regards the **number of animals investigated for non(neoplastic) lesions** at the end of the study, the OECD GD 116 specifies that not only control and high dose animals are to be investigated, but also all potential target organs in mid dose animals, plus animals deceased before scheduled death or sacrificed due to moribund status. Based upon this, the probability to overlook a finding was considered as very low.

For **findings with high background incidence** (e.g. mammary gland fibroadenoma), HCD may be particularly important to identify and interpret low, aberrant incidence in the concurrent control comparing to treatment groups. Although such unusually low incidence in the concurrent control could question the validity of the study, one alternative could be to accept the study in a whole weight of evidence approach, including knowledge from HCD.

Intense discussion on **hormonal (especially thyroid hormones) measurement** revealed that this is very sensitive to different variables (time point for blood collection, equipment, method of anaesthesia, age of animals, etc.) and therefore a too sensitive parameter to compare the results between different laboratories. While TSH and histopathological findings in thyroid are closely correlated, reduced T4 need not be correlated to any histopathological finding in order to be classified as adverse in a reproductive toxicity study, since its decrease in dams can cause severe effects in development of offspring. The level of adversity for decrease in T4 (20%, e.g.) was discussed but it was recognised that this is outside the scope of the workshop topics.

Difficulties in having pre-treatment values for hormones in offspring, differently than for dams, has been highlighted. In this case information from HCD can be important to understand the finding, in addition to the concurrent control.

It was reiterated that **high level of granularity** is needed when HCD are submitted, in order to assess their comparability to concurrent control. This high level of granularity includes more details on study conditions, availability of metadata sufficiently accurate, as well as in many cases individual study data (rather than summary statistics), which can be used for an independent assessment by the regulators.

The possibility to extend the **5 years time-period** for acceptance of HCD strongly depends upon the availability/number of data and the stability in the response level of the endpoint. The discussion reflected the dilemma of short time period (e.g. 5 years) *versus* sufficient number of studies/ data points to have a robust database. This might be especially controversial for studies of longer duration (e.g. carcinogenicity studies) or with low incidence results (e.g. rare malformations) where limited number of data points is available for short time period. In addition, it may depend upon the occurrence of a finding in the control group and be rather justified where concurrent control shows a borderline increase. If **merging of studies** with e.g. different administration route is intended in order to increase the number of data points, it should be shown that the administration route has no impact on the investigated parameter. Well-justified combination of studies in order to increase the benefit of already available data is also relevant in context of 3R principles (Replacement, Reduction and Refinement) as one of the utmost goals in the regulatory toxicology.

It was reiterated that in case of extending the relevant time-period for HCD, it has to be shown that there is no trend for changes in the respective parameter.

It was briefly touched on which date, as mentioned in the study (start of the study, end of in life phase), should be the **starting point for consideration of relevant time period** (for concurrent study and HCD). This information should be clearly presented when are submitted.

When a **new parameter/marker** is introduced in the investigation, no HCD are available. Even if data from validation assays or vendor data might be considered to some extent, it has to be proven that study conditions were comparable.

As regards use and **interpretation of HCD in genetic toxicology** it was emphasized that HCD are not used in every assay as a tool for deciding upon the biological relevance of the finding (e.g. Mouse lymphoma assay, Ames assay).

Inclusion of HCD in an assessment means that **more variability** is included than represented in the control group population, but on the other hand, **the real population** is potentially more broadly covered. It was questioned if HCD, once considered acceptable, should be used for interpretation of every endpoint in the relevant study, not only for interpretation of those endpoints for which treatment-related and adverse effects are discussed. In that case, much more data would need to be evaluated and probably full studies from HCD pool to be submitted. This is however not only a question of time, but also of proprietary rights of study owners.

As regards **ecotoxicity studies** (including avian reproductive studies) it was recognised that the principles for use and reporting of HCD are basically the same as for toxicity studies, but the interpretation of results can be specific for ecotoxicity purpose (population versus individual). It was recognised that HCD could be important for ecotoxicity studies as well, to put the control response into perspective.

Importance of a Guidance for harmonisation of criteria on use and reporting of HCD within and between different EU agencies (e.g. EFSA, ECHA) was highlighted, in order to increase the understanding and trust in the assessment by third parties not deeply involved in the process. For this purpose, also clear requirements for data compilation should be communicated to the contract research organisations.

3. Feedback to the workshop

Following the workshop, all participants (182 in total) were kindly asked to respond (anonymously) in the following questions:

Question A

Based on the discussions held during the workshop, please specify three topics of high priority for further considerations. (Please include just 3 keywords, if possible) *1. …..* *2. …..* *3. …..*

Question B

Do you think that the workshop program and the discussions held captured all relevant issues? (YES/NO) *If 'NO' what do you think that was missing?*

Forty-four (44) responses have been provided in total.

The majority of the responders highlighted the need for harmonisation and the development of specific data requirements and guidance on the use and interpretation of HCD, including data requirements and reporting (format). In addition, reference has been made to relevance/suitability or/and reliability/acceptability criteria, while regarding the prioritization of more specific issues to be dealt with the statistics and the relevant time period to be considered have also been pointed out. Although almost one third of the responders mentioned a specific toxicity area (e.g. genotoxicity of carcinogenicity) as a priority, there has been no consensus in the area to be prioritised.

Forty (40) participants (91% of the responders) answered that the workshop program and the discussions held captured all relevant issues. Those stating that there were topics missing referred more to concerns regarding the use of HCD and not to a particular issue that was not discussed within the workshop.

4. Conclusions

The intense discussion during the workshop revealed the need for:

- Clear set of criteria to be fulfilled for acceptability of HCD. These criteria can depend upon the relevant endpoint
- High granularity in presentation of HCD and their metadata to recognise differences between control populations (HCD and contemporaneous study) coming from intra- and inter studies variation
- Close exchange between different domains (statistics, toxicology) before appropriate statistics is applied for consideration of HCD

5. Recommendations

Based upon the discussion in the workshop following recommendations emerged:

- To develop a set of criteria/guidance for use of historical control data in toxicity studies.
- To involve experts from different domains when criteria/guidance is developed
- To harmonise the use, reporting and interpretation of historical control data within and between national regulatory bodies and European Agencies (EFSA, ECHA)

References

BPR Regulation 2012. Regulation (EU) No 528/2012 of the European Parliament and of the Council of 22 May 2012 concerning the making available on the market and use of biocidal products.

Guidance on the application of the CLP criteria, 2017. Guidance to Regulation (EC) No 1272/2008 on classification, labelling and packaging (CLP) of substances and mixtures. Version 5.0. July 2017

JMPR 2015. Guidance document for WHO monographers and reviewers. WHO 2015

Abbreviations

CLH	Harmonised classification and labelling
DART	Developmental and reproductive toxicity studies
ESTP	European Society of Toxicologic Pathology
goRENI	Global open Registry Nomenclature Information System
IGS	International Genetic Standardization
IFTS	International Federation of Teratology Societies
INHAND	International Harmonization of Nomenclature and Diagnostic criteria
NTP	National Toxicology Programme
REML	Residual Maximum Likelihood
RI	Reference intervals
RITA	Registry of Industrial Toxicology Animal-data
SOP	Standard operating procedures
T4	Thyroxin
TSH	Thyroid stimulating hormone

Annex A - Programme of the workshop

Annex B - Presentations

Annex C – Feedback to the workshop

Annex D - List of participants