Food for Thought …

Introducing the Concept of Virtual Control Groups into Preclinical Toxicology Animal Testing

Thomas Steger-Hartmann¹, Annika Kreuchwig¹, Lea Vaas¹, Jörg Wichard¹, Frank Bringezu², Alexander Amberg³, Wolfgang Muster⁴, Francois Pognan⁵, Chris Barber⁶

¹Bayer AG, Pharmaceuticals, Investigational Toxicology, Berlin, Germany; ²Merck Healthcare KGaA, Biopharma and Non-Clinical Safety, Darmstadt, Germany; ³Sanofi, Preclinical Safety, Frankfurt, Germany; ⁴Roche Pharmaceutical Research & Early Development, Pharmaceutical Sciences, Roche Innovation Center, Basel, Switzerland; ⁵Novartis Institute for Biomedical Research, Basel Switzerland; ⁶Lhasa Ltd. Leeds, UK

Abstract
Sharing legacy data from in vivo toxicity studies offers the opportunity to analyze the variability of control groups stratified for strain, age, duration of study, vehicle and other experimental conditions. Historical animal control group data may lead to a repository, which could be used to construct virtual control groups (VCGs) for toxicity studies. VCGs are an established concept in clinical trials, but the idea of replacing living beings with virtual data sets has so far not been introduced into the design of regulatory animal studies. The use of VCGs has the potential of a 25% reduction in animal use by replacing the control group animals with existing randomized data sets. Prerequisites for such an approach are the availability of large and well-structured control data sets as well as thorough statistical evaluations. the foundation of data sharing has been laid within the Innovative Medicines Initiatives projects eTOX and eTRANSAFE.

For a proof of principle participating companies have started to collect control group data for subacute (4-week) GLP studies with Wistar rats (the strain preferentially used in Europe) and are characterizing these data for its variability. In a second step, the control group data will be shared among the companies and cross-company variability will be investigated. In a third step, a set of studies will be analyzed to assess whether the use of VCG data would have influenced the outcome of the study compared to the real control group.

1 Introduction

Virtual control groups (VCGs), also called synthetic control arms, represent a concept, which is well established for randomized clinical trials (Berry et al., 2017). VCGs are used to compare an experimental treatment with the standard of care. The main purpose of using virtual participants or patients instead of true patients is to increase the speed of the trials while avoiding costly recruitment of probands, thus allowing all recruited patients to receive the new treatment (ethical considerations). The concept is particularly used in clinical cancer studies and assumes that if the range of variables are kept sufficiently constant between the patients foreseen for a new experimental treatment and those who have received the standard of care, the outcome of the clinical trial can be assessed based on the comparison between the virtual control arm and the patients treated with the new experimental design (Switchenko et al., 2019). In the clinical setting, such virtual control arms can be constructed in several ways. The simplest approach is the use of historical control data for the respective endpoint under examination (Berry et al., 2017), which is also described in the guideline of the International Council of Harmonisation (ICH) E10 (ICH, 2001). However, differences in the patient selection can easily confound the findings in the comparison of new treatments with historical controls. To overcome this problem, VCGs can also be constructed based on the clinical characteristics of the chosen patients using statistical models such as Bayesian approaches (Spiegelhalter, 2004) or nomograms which for example estimate progression-free survival under standard of care in oncology trials (Jia et al., 2014). A third approach of constructing VCGs is based on the use of Electronic Health Records (EHR), i.e. the systematized collection of patient and population electronically-stored health information in a digital format (Eichler et al., 2016).

Although the use of historical data is well established in preclinical studies for comparative purposes (Yanagawa and Hoel, 1985; Haseman, 1992), the concept of VCGs to actually replace control group animals has not yet found its entry in regulatory toxicity testing because compared to recruitment of patients, the availability of control animals is not a technical issue and in comparison to clinical studies, the cost and time factor do not play as an important role in the conduct of animal studies. However, ethical consideration, i.e. the strive to reduce animal numbers in preclinical research has triggered consideration of reducing or replacing control groups. In the area of in vivo genotoxicity study Pfulker et al. (2009) investigated possibilities and acceptance of reducing the size of positive control groups or completely omitting concurrent positive control
groups with reference to historical positive control data. For pharmacological studies, Kramer and Font (2016) have proposed strategies for reducing control group sizes in animal studies by incorporating historical control group data. Despite these efforts, replacing control groups in systemic toxicity studies has not yet been approached.

The conventional setting of a regulatory toxicology study uses 25% of the animals for controls: OECD guideline TG 407 “Repeated Dose 28-Day Oral Toxicity Study in Rodents” (OECD, 2008a) may be quoted here as an example for subacute studies, though the pertinent guidelines for chronic studies or carcinogenicity studies essentially follow the same schema (OECD, 2008b) with the exception that the animal numbers per group increase with the duration of the study. The quoted guideline requires at least three dose groups and one control group. Animal numbers in these four-week rodent studies are normally 10 per dose group or control (five females and five males) and may increase if interim euthanasia or recovery groups with satellite animals are planned. Four-week rat studies for food ingredients or pharmaceutical compounds usually apply higher animal numbers which amount to 10 per sex and dose (FDA, 2003a). Numbers may even be higher if recovery groups are included in the study design which than usually also have parallel control groups. For non-rodents, three to four animals per sex and dose groups are used in four-week studies (FDA, 2003b). In chronic studies, animal numbers for rodents might increase up to 50 rats per sex and dose group and 10 dogs or primates per sex and dose group (Gad, 1995).

2 Current use of historical control data in preclinical animal studies

None of the above-mentioned guidelines foresee the use of VCGs, though it is advised that “historical control data are collected and that for numerical data, coefficients of variation are calculated” (OECD, 2008a). The main purpose of this data collection is the performance control of the study and the assessment of outliers which may occur in individual studies for various reasons. Legacy data from control animals are used for determining the range of parameters of untreated animals, its changes over time or the influences of changes in analytical methods.

Petterino and Argentino-Storino (2006) list three applications of historical control data in the area of clinical pathology and hematology:

1. Evaluation of the clinical and hematological pathology data
2. Evaluation of background pathology in animal populations
3. Comparison of analytical methods

If a statistically significant difference between a dose group and the control group is observed in a study for a specific parameter, but the changes in the treated group lie within historical control ranges, then it is questionable whether the observation actually represents a compound-related effect.

Historical control data are of particular interest for the evaluation of carcinogenicity studies with respect to incidences of spontaneous tumors observed, which depend on species and the used strain (Morawietz et al., 1992). In addition to the direct comparison with the control group such a comparison with historical control data allows the assessment whether the occurrence of a rare tumor or a marginally increased tumor incidence is of biological relevance, i.e. caused by the chemical under investigation (Yanagawa and Hoel, 1985; Haseman, 1992; Greim et al., 2003). For the assessment of developmental toxicity studies, the situation regarding historical control data is similar compared to carcinogenicity studies. The data collection on fetal findings, including incidences of spontaneous external, visceral, and skeletal anomalies contributes to the differentiation between spontaneous findings or true toxic effects particularly for rare findings (Kuwagata et al., 2019).

In summary, the value of historical control data has been referenced for decades, however, a replacement of control group animals with virtual animals based on existing data has been proposed only recently for animal disease models with surgical interventions (Kramer and Font, 2016; 2017). The concept has, however, not found broad repercussions in regulatory safety testing probably also because of the lack of appropriate large and easily accessible control animal data sets.

3 Challenges for the use of Virtual Control Groups

There are several reasons why existing data from control group animals cannot be easily used for constructing virtual control groups. The first reason, which was broadly discussed in a publication on historical data use for assessing carcinogenicity data (Greim et al., 2003), is the heterogeneity of the data sets. The physiology of experimental animals and thus the data acquired during animal studies is influenced by numerous factors such as genetic disposition, strain, animal housing conditions, diet, stress during handling and administration, age of the animals, and infections. Some of these factors are under the influence of the investigator and thus can be controlled and are also defined by the guidelines or related good practice documents. Such controllable factors are strain, age, diet, and housing conditions (temperature, lighting, humidity, bedding, single or group housing). Others are less amenable to standardization such as stress during animal handling or infections occurring during the course of the study. An important aspect, which can only be controlled by stringent breeding programs (Low-Marchelli 2017) is genetic drift in the inbred animals used in toxicity studies. Genetic drift may remain silent and go undetected until a new phenotype is identified by the analytics used in the studies, including histopathology. However, it represents a cause for irreproducibility of experimental results (Brekke et al. 2018). As long as control animals are used, these underlie the same level of genetic drift. A replacement by virtual controls using historical data might result in differences between the treatment groups and the virtual control groups. These might falsely be attributed to a test compound effect though they are rather caused by genetic differences between the previously used animals and those in the current study.

A further reason for using control groups is the assessment of infections which may influence hematological and histopathological parameters of the affected animals (GV-SOLAS, 1999). Without a control group such spontaneous infections might erroneously be attributed to effects elicited by a test compound. In a more general sense, the use of control groups in this context also allows to assess the interaction between animal behavior and the laboratory environment, which was identified to represent a major contribution to variability of results between test sites (Crabbe et al. 1999).
The third reason is the interlaboratory and time dependent variability of analytical methods used for clinical chemistry evaluations, which constantly undergo changes due to improvements of the analytical technologies applied (e.g. reduced sample volumes, improved detection limits, changed assay composition). These are usually tracked within the laboratory by using quality control samples and also evaluated by ring tests. However, such changes limit the applicability of control group data to certain time periods for which identical methods and biochemical assays (“kits”) have been used.

A fourth reason relates to non-numerical, descriptive findings gathered during animal studies such as gross pathology and histopathology. In order to make these findings interoperable, comparable and compatible between studies, there is a need for controlled vocabularies and ontologies (Hardy et al. 2012). Despite tremendous efforts by the Societies of Toxicologic Pathology to standardize the vocabularies and terminologies through the International Harmonization of Nomenclature and Diagnostic Criteria (INHAND) (Creasy et al., 2012), the heterogeneity of terms for describing similar or identical findings is still an issue. This was also addressed in the IMI eTOX project by setting up ontologies and controlled terminologies for histopathology (Briggs et al., 2015). Standardization and harmonization of terminologies is of particular importance, since in contrast to numerical parameters the original observation in an animal study, i.e. the microscopic slide, is usually not stored as image in databases, thus preventing a rapid automated re-assessment. Complex guidelines have been developed for histopathology peer review for toxicology studies (OECD, 2014) to assure the quality and the accuracy of interpretation, but these regulations do not solve the issue of varying terminologies. The implementation of the Standard for Exchange of Non-clinical Studies (SEND) developed by Consortium of Standards in the Clinical Research Process (CDISC. 2016) has triggered further harmonization of terminologies for toxicology studies in the pharmaceutical industry. This standard is mandatory for repeated-dose toxicity studies used in new drug applications (NDAs) at FDA since December 2016 (FDA, 2014) and is gradually extended to other types of toxicity studies. The fifth and probably most important aspect is the lack of appropriate large data sets and repositories for control animal data. Study centers, including contract research organizations (CROs) usually set up their own inventories for the collection of historical data, but these data sets are not broadly shared, except for above mentioned purposes in the assessment of incidences of spontaneous neoplasms (Registry of Industrial Toxicology Animal data [RTA] or the North American Control Animal Database [NACAD]) (Morawietz et al., 1992; Greim et al., 2003) or developmental findings. Most of the reference data sets for physiological parameters are published in the literature (Loeb and Quimby, 1989) or in disparate databases with rather limited query functionalities thus limiting their usability.

4 Steps towards the implementation of VCGs

4.1 Data and data repository requirements

Some of the hurdles mentioned above have already been addressed in the context of historical data comparisons. Greim et al. (2003) lists requirements, which need to be fulfilled by historical data sets in order to be eligible for comparative re-use in the context of carcinogenicity assessments:

1) The historical data were obtained with the same species and strain of experimental animal. The animals were acquired from one supplier only.
2) The historical control data were produced in the same laboratory as the experimental data.
3) The study design, experimental methods and assessment criteria were the same (i.e. parameters such as age of the animals at the beginning of the experiment, animal housing conditions, methods of obtaining samples (e.g., number, size, localization and orientation of the tissue sections), diagnostic criteria (e.g., standardized terminology, peer review of critical findings by expert pathologists) need to be assessed and controlled for.
4) The studies used as historical controls were carried out during a defined time window to limit the variability of the parameters.

With exception of requirement No. 2, these criteria can also be applied to the VCG approach. The idea of VCGs is based on large shared data sets from toxicology studies for regulatory dossiers across industry. The shared data should have the highest possible level of granularity (individual animal data for all parameters acquired during a study plus meta information on the conditions how the data was measured and assessed) to allow analyses, how far certain data sets are comparable across studies, laboratories and for specific time periods.

Such a control animal data collection could thus not only be used to reduce animal numbers in control groups, but also to invert some of Greim’s criteria by applying rigorous statistical procedures: if a data set from an individual animal falls within pre-defined statistical limits, it can be used for a VCG, independently whether the data were produced in the same laboratory or with the same procedure, hereby partly overcoming the described problem of statistically “under-powered” animal studies (Hartung, 2007).

Prerequisite for the concept of VCGs is the availability of a repository for shared data, which is accessible not only to companies providing data, but also institutions assessing the data, i.e. regulatory authorities or the scientific community. The infrastructure and the legal framework for sharing toxicity data has already been established during the recent IMI project eTOX (Sanz et al., 2017). Since the SEND implementation with its standardized terminologies came into force just towards the end of this project, the follow-up IMI project eTRANSAFE started in 2017 now focusses on data sharing using these new standards. eTRANSAFE is a five-year project, funded by the IMI-2 Joint Undertaking together with the pharmaceutical industry, that aims to develop an advanced data integration infrastructure together with innovative computational methods to improve the safety in drug development processes. Key component of the data sharing activities is the so-called honest broker, who receives and protects the data from the participants, and generates as well as manages the eTRANSAFE.

1 www.etransafe.eu
data warehouse. The honest broker also administers the access rights to the data base and ensures the separation of confidential from non-confidential shared data. This separation is of importance since the complete shared study may contain data which is considered as sensitive information by the providing company particularly with regard to compound information (e.g., structure, target or indication), whereas the control group animal data requires no intellectual property protection, i.e., can be shared freely among all participating partners and potentially also with future users not belonging to the project, including regulators or the scientific community.

It is important to keep in mind that for the ultimate implementation of the VCG concept the data repository needs to fulfill certain criteria for computer system validation. Regulatory animal studies are performed under Good Laboratory Practices (GLP). Therefore, the use of historical control data for constructing the VCG underlies the same principles, namely the requirements set forth in the advisory document “Application of GLP Principles to Computerised Systems” (OECD, 2016).

4.2 Statistical procedures
The feasibility of the proposed concept will be assessed by the participants of the eTRANSAFE project in a stepwise approach. Initially, the companies will analyze their internal historical control repositories and subject these to internal statistical analyses. Key components are descriptive statistics on data distribution and general characteristics like those which have previously been published for the inclusion for historical control data for the micronucleus assay (Igl et al., 2019). Purpose of these statistical consideration and the stratification of data are to gain insight into general features such as variability of parameters and how these change over time. This will be visualized with the help of control charts. Key factors of experimental conditions which influence the variability shall be identified and numerically described. Such factors could be for example changes in analytical procedures for measuring clinical chemistry parameters or modifications in the housing conditions. A list of the minimum parameters to be collected and analyzed is given in Table 1.

Tab. 1: List of minimum parameters to be collected and analyzed in the context of the VCG concept

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Strain, Supplier</td>
<td>Body weight &amp; weight gain</td>
</tr>
<tr>
<td>Age</td>
<td>Clinical observations and behavior (in-life observations)</td>
</tr>
<tr>
<td>Sex</td>
<td>Clinical chemistry</td>
</tr>
<tr>
<td>Weight</td>
<td>Hematology</td>
</tr>
<tr>
<td>Study start</td>
<td>Coagulation</td>
</tr>
<tr>
<td>Study duration</td>
<td>Organ weights (absolute and relative)</td>
</tr>
<tr>
<td>Route of administration</td>
<td>Gross pathology</td>
</tr>
<tr>
<td>Group size</td>
<td>Histopathology</td>
</tr>
<tr>
<td>Vehicle information</td>
<td>Food &amp; water consumption</td>
</tr>
<tr>
<td>Body weight &amp; weight gain</td>
<td></td>
</tr>
<tr>
<td>Clinical observations and behavior (in-life observations)</td>
<td></td>
</tr>
<tr>
<td>Clinical chemistry</td>
<td></td>
</tr>
<tr>
<td>Hematology</td>
<td></td>
</tr>
<tr>
<td>Coagulation</td>
<td></td>
</tr>
<tr>
<td>Organ weights (absolute and relative)</td>
<td></td>
</tr>
<tr>
<td>Gross pathology</td>
<td></td>
</tr>
<tr>
<td>Histopathology</td>
<td></td>
</tr>
<tr>
<td>Food &amp; water consumption</td>
<td></td>
</tr>
</tbody>
</table>

If these factors are sufficiently characterized, the proof of principle would use the collected control group data to construct VCGs for studies which have been performed in the past. Despite Bayesian approaches, two options are possible: 1) the “resampling approach”: the VCG data shall be drawn in a randomized way out of a pool of animal data sets, which matches the actual control group of the study regarding the above-mentioned key factors. The studies shall be re-analyzed by using the VCGs instead of actually used control animals applying the accepted procedures of choice. 2) The second approach is called a “simulation approach”: the gained information about the distribution of endpoints are used to simulate the necessary amount of virtual control data and thus provide virtual control data. With those newly derived control data the study shall be reanalyzed analogous to approach No. 1. Figure 1 illustrates the described steps for replacing control animals by virtual control groups.

Fig. 1: Graphical illustrations of the five steps in the replacement of animals in control groups by VCGs
In cases where differences in the assessment of compound-related toxicities are detected due to the use of the VCGs, a root cause analysis will be performed to identify, which variation of parameters cause the deviating assessment. These analyses will be performed first within individual companies but subsequently also across companies in order to investigate the level of variability across companies or test facilities. Assessing the shared and pooled data is important since with larger data sets from different sources the distribution of normal ranges will naturally broaden. This needs to be strictly controlled in order to be able to delineate normal findings from treatment-related findings in these larger control animal data sets.

4.3 Steps towards scientific and regulatory acceptance

The eTRANSAFE project has a scientific advisory board consisting of members with significant regulatory background. In a first step, these members will assess the feasibility of the VCG concept. If the described analyses allow a proof of concept, it is intended to approach EMA within its Innovation Task Force (EMA, 2014) to present the approach and explore its regulatory acceptance.

We expect that the acceptance of the VCG concept will be achieved step-wise, where initially the size of control groups will be reduced, and the omitted animals complemented with VCG data. In a subsequent step, after gaining sufficient experience and confidence for allow for a complete replacement of control animals, so-called sentinel animals can be housed within the animal facilities (Lipman, 2003). These sentinel animals can be used for several studies and only need to be investigated in case of suspected infections or other expected laboratory environment influences, thus overcoming the respective concerns.

5 Conclusion

Even though regulatory acceptance across the different region (EMA, FDA, JMMA) will probably take years, if not decades, the data collection is nevertheless an important asset and advancement for the 3Rs concept. The EMA guideline on repeated toxicity study (EMA, 2008) states that the size of treatment groups is inter alia dependent on the “background knowledge concerning the ranges of variables to be studied in the species and strains”.

Assessing this variability of control group data through the availability of large data sets offers the chance to significantly reduce the group sizes of control groups, hopefully even omitting them. Improving the background knowledge will also enhance the assessment whether an observation is treatment or compound-related or rather a spontaneous finding. An improved understanding of within- and between-animal-variation will furthermore be helpful in reducing the uncertainty in the determination of no (adverse) effect levels (NO(A)EL) (Paparella et al. 2013). It can also contribute to select the most appropriate strain of species for specific studies e.g. for carcinogenicity studies, where spontaneous tumor rates differ between strains (Greaves and Rabemannianina, 1982, Morawietz et al. 1992).

In conclusion, the success of the VCG concept will enable a significant reduction of animal use in repeated dose toxicity studies, thus substantially contributing to the 3R concept.

References


**Conflict of Interest**
The authors of this publication are employees of the participating companies or organizations. The contribution to this publication was part of their work and is paid by their salaries.

**Acknowledgements**
The described initiative is performed under the IMI eTRANSAFE project. eTRANSAFE has received support from IMI2 Joint Undertaking under Grant Agreement No. 777365. This Joint Undertaking receives support from the European Union’s Horizon 2020 research and innovation program and the European Federation of Pharmaceutical Industries and Associations (EFPIA).