"I am not an advocate for frequent changes in laws and constitutions, but laws and institutions must go hand in hand with the progress of the human mind. As that becomes more developed, more enlightened, as new discoveries are made, new truths discovered and manners and opinions change, with the change of circumstances, institutions must advance also to keep pace with the times. We might as well require a man to wear still the coat which fitted him when a boy as civilized society remain ever under the regimen of their barbarous ancestors."

-Thomas Jefferson, 1816

**t4 Workshop Report**

**Toward Implementing Virtual Control Groups in Nonclinical Safety Studies: Workshop Report and Roadmap to Implementation**

Emily Golden1, David Allen2#, Alexander Amberg3, Lennart T. Anger4, Elizabeth Baker5, Szczepan W. Baran6#, Frank Bringezu7#, Matthew Clark8, Guillelmette Duchateau-Nguyen9#, Sylvia E. Escher10#, Varun Giri11#, Armelle Grevot12, Thomas Hartung13,14, Dingzhou Li14, Laura Lotfi15, Wolfgang Muster9#, Kevin Snyder16, Ronald Wange16, Thomas Steger-Hartmann17#

1Center for Alternatives to Animal Testing (CAAT), Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA; 2Inotiv, Inc., Morrisville, NC, USA; 3Sanofi, Preclinical Safety, Frankfurt, Germany; 4Genentech Inc., South San Francisco, CA, USA; 5Physicians Committee for Responsible Medicine, Washington, DC, USA; 6VeriSim Life, San Francisco, CA, USA; 7Merck Healthcare KGaA, Chemical & Preclinical Safety, Darmstadt, Germany; 8Charles River Laboratories, now KALOS Technologies, Philadelphia, PA, USA; 9Roche Pharmaceutical Research & Early Development, Pharmaceutical Sciences, Roche Innovation Center, Basel, Switzerland; 10Fraunhofer ITEM, Hannover, Germany; 11BASF SE, Experimental Toxicology and Ecology, Ludwigshafen am Rhein, Germany; 12Novartis Institute for Biomedical Research, Novartis AG, Basel, Switzerland; 13CAAT-Europe, University of Konstanz, Konstanz, Germany; 14Pfizer, Global Biometrics and Data Management, Groton, CT, USA; 15Charles River Laboratories, Boston, MA, USA; 16Food and Drug Administration, Center for Drug Evaluation and Research, Silver Spring, MD, USA; 17#Bayer AG, Pharmaceuticals, Investigational Toxicology, Berlin, Germany

**Abstract**

Historical data from control groups in animal toxicity studies is currently mainly used for comparative purposes to assess validity and robustness of study results. Due to the highly controlled environment in which the studies are performed and the homogeneity of the animal collectives it has been proposed to use the historical data for building so-called virtual control groups, which could replace partly or entirely the concurrent control. This would constitute a substantial contribution to the reduction of animal use in safety studies. Before the concept can be implemented, the prerequisites regarding data collection, curation and statistical evaluation together with a validation strategy need to be identified to avoid any impairment of the study outcome and subsequent consequences for human risk assessment. To further assess and develop the concept of virtual control groups the transatlantic think tank for toxicology (t4) sponsored a workshop with stakeholders from the pharmaceutical and chemical industry, academia, FDA, pharmaceutical, contract research organizations (CROs), and non-governmental organizations in Washington, which took place in March 2023. This report summarizes the current efforts of a European initiative to share, collect and curate animal control data in a centralized database and the first approaches to identify optimal matching criteria between virtual controls and the

**6Organizing Committee**

*A report of t4 – the transatlantic think tank for toxicology, a collaboration of the toxicologically oriented chairs in Baltimore, Konstanz and Utrecht sponsored by the Doerenkamp-Zbinden Foundation. The spectrum of views expressed in this article are those of the contributing authors and do not necessarily reflect those of their institution of employment.

Received October 4, 2023; Accepted November 22, 2023; Epub December 1, 2023; © The Authors, 2023.

**ALTEX 41(4), ###-###. doi:10.14573/altex.2310041**

Correspondence: Thomas Steger-Hartmann, PhD
Bayer AG, Pharmaceuticals, Investigational Toxicology, Mülnerstraße 178, 13353 Berlin, Germany (thomas.steger-hartmann@bayer.com)
treatment arms of a study as well as first reflections about strategies for a qualification procedure and potential pitfalls of the concept.

Plain language summary
Animal safety studies are usually performed with three groups of animals where increasing amounts of the test chemical are given to the animals and one control group where the animals do not receive the test chemical. The design of such studies, the characteristics of the animals, and the measured parameters are often very similar from study to study. Therefore, it has been suggested that measurement data from the control groups could be reused from study to study to lower the total number of animals per study. This could reduce animal use by up to 25% for such standardized studies. A workshop was held to discuss the pros and cons of such a concept and what would have to be done to implement it without threatening the reliability of the study outcome or the resulting human risk assessment.

1 Background

Virtual control groups (VCGs) based on electronic health records are frequently implemented in clinical trials (Berry et al., 2017), where they are used to compare existing treatment protocols to experimental therapies. However, the concept of VCGs has yet to take root in regulatory toxicity testing. In 2020, Steger-Hartmann et al. introduced the concept of VCGs for use in nonclinical testing, where a VCG, which is constructed using adequately selected historical control data, replaces, partially or completely, the concurrent control (CCG) in a nonclinical safety study (Steger-Hartmann et al., 2020). Using VCGs in nonclinical safety testing could result in a reduction in animal use by up to 25% per study (Steger-Hartmann and Clark, 2023). This approach is fundamentally in line with the shifts occurring both scientifically and societally, as calls for reduction or replacement of animal tests, whenever scientifically appropriate, have become mainstream (DHHS, 2023; US FDA, 2022; NASEM, 2023; NIH, 2022).

Moreover, VCGs have the potential to relieve the pressures created by the shortage of non-rodent laboratory animals, especially NHPs, in the field of biomedical research. Although only a relatively small number of NHPs are used in research compared to other species (e.g., rodents), they are often the non-rodent species used in the drug approval process, particularly for biologics (Phillips et al., 2014). This shortage creates not only major financial problems, as this supply chain bottleneck exponentially increases costs, but also creates a public health problem, as supply chain delays preclude promising drug candidates from reaching clinical trials, thereby delaying the drug approval process. A recent report from the National Academy of Sciences also concluded that the NHP shortage has continued to worsen (NASEM, 2023). For these reasons as well as ethical concerns, drug regulators and drug developers should think about reducing animal numbers, especially NHPs, through VCG implementation wherever possible. Mecklenburg et al. (2023) recently published a study where they evaluated twenty subchronic NHP toxicity studies to determine if the hazards reported in the original study could have also been correctly identified without reference to CCGs. The authors concluded that toxicological hazards of key relevance were identifiable without reference to the CCGs mainly through the alternative use of pre-dosing values as reference. The authors therefore propose that this study type represents a good starting point to implement VCGs to reduce NHP usage (Mecklenburg et al., 2023).

To this end, the transatlantic think tank for toxicology (t4) sponsored a workshop to discuss the feasibility of VCGs in nonclinical safety testing. The 3-day workshop was held in Washington, DC in March 2023 and included representatives from the pharmaceutical and chemical industry, academia, FDA, pharmaceutical, contract research organizations (CROs), and non-governmental organizations (NGOs). During the workshop, the VCG concept for nonclinical testing was introduced, and several facets crucial to the success of VCGs in nonclinical testing, including creation of a VCG database, statistics and selection criteria, and steps towards regulatory acceptance, were presented. Following these presentations, focus sessions addressed remaining data gaps and defined the statistical procedures necessary to generate VCGs. Finally, a roadmap to move the concept forward was outlined that identified next steps and the main issues to be addressed before VCGs can become reality. The results of the workshop are summarized in this report.

In addition to the listed authors, the following individuals were also attendees at the workshop: Todd Bourcier, FDA, USA; Peter Brinck, Novo Nordisk, Denmark; Susan Butler, ORISE Fellow, FDA, USA; Mark Carfagna, Eli Lilly, USA; Warren Casey, National Institutes of Health, Division of the National Toxicology Program, National Institutes of Environmental Health Sciences, USA; Karen Davis Bruno, FDA, USA; Suzanne Fitzpatrick, FDA, USA; Andrew Goodwin, FDA, USA; Sven Kronenberg, Roche Pharmaceutical Research & Early Development, Pharmaceutical Sciences, Roche Innovation Center, Basel, Switzerland; Elisa Passini, NC3Rs, UK; David Potter, Pfizer, USA; Md Aminul Islam Prodhann (Amin), ORISE Fellow, FDA, USA; Md Yousuf Ali, ORISE Fellow, FDA, USA; Haleh Saber, FDA, USA. Nicole Kleinstreuer (NICEATM) was a member of the organizing committee but was unable to attend the workshop.

Abbreviations: AI, artificial intelligence; BW, body weight domain in SEND; CL, clinical observations domain in SEND; CSV, computer systems validation; DM, demographics domain in SEND; GLP, good laboratory practice; LB, laboratory domain in SEND; MA, macroscopic findings domain in SEND; MI, microscopic findings domain in SEND; ML, machine learning; OM, organ measurements domain in SEND; QA, quality assurance; SEND, standard for exchange of nonclinical data; TS, trial summary domain in SEND; VCG, virtual control group; VICOG WG, Virtual Control Group Working Group
The workshop was opened with a brief history on the concept of VCGs. The origin of the VCG concept came from an Innovative Medicine Initiative (IMI) project called eTOX (electronic toxicity), where aggregated toxicity data from pharmaceutical companies were shared. The purpose of this database was to take advantage of the substantial amount of nonclinical toxicity data produced by the pharmaceutical industry to potentially improve efficiency of drug development while at the same time reducing the number of animals used in testing (Briggs et al., 2012). This ultimately led to a larger project called eTRANSAFE (Enhancing TRANslational SAFety Assessment through Integrative Knowledge Management), where the overarching goal was to translate the data to humans (Pognan et al., 2021; Sanz et al., 2023). Initially, there were concerns regarding data sharing; however, the companies overcame these concerns by only providing the data that lacked any intellectual property (IP) sensitivities, such as control group data without any compound or biological target identifiers. While this was useful to calculate normal reference values on large sample sizes and background incidences of pathological findings, the question was raised as to whether more could be done with these data. Here, the concept of VCGs for nonclinical safety was born.

The concept of VCGs in nonclinical testing is that historical control data are collected in a central database, and using appropriate stratification and randomized selection, a VCG can be generated to augment/replace CCGs in nonclinical toxicity tests. A more detailed introduction to this concept is outlined elsewhere (Steger-Hartmann et al., 2020; Steger-Hartmann et al., 2023). The eTRANSAFE project ended in February 2023, but some partners continued to explore the data for the purpose of VCG. A VCG working group, formally titled the Virtual Control Group Working Group (VICOG WG), was established, and, at present, five European Federation of Pharmaceutical Industries and Associations (EFPIA) eTRANSAFE partners (Bayer, Merck Healthcare KGaA, Sanofi, Roche, and Novartis) and two public partners (UPF/IMI and FhG-ITEM) are involved. The three main tasks identified by the working group to get VCGs off the ground were to: 1) establish a curated and shared database (VICOG); 2) develop matching procedures using statistical tools for data selection and visualization and optimize these; and 3) establish a qualification procedure.

3 Proof-of-Concept: Evaluation of the Initial VCG Framework

To gain understanding of the practicality of the VCG approach, members of the VICOG working group evaluated the proposed VCG framework and presented the results of first proof-of-concept studies.

3.1 VICOG Database Collection and Curation

The foundation for putting VCGs into practice is ensuring a high-quality database from which VCGs can be generated. A summary of the current prototypical database was presented, termed the VICOG database. It is constructed from data provided by the
Fig. 2: Overview of the data collection for the VICOG database up to March 2023

aforementioned EFPIA partners, including legacy study data and current studies in the Standard for Exchange of Nonclinical Data (SEND) format, as summarized in Figure 1.

Following extraction of SEND data, data are harmonized via data preparation processes, including transformation of legacy study data into SEND terminology, and are subjected to an “enrichment” process using the trial summary. The trial summary provides highly variable information depending on the study. Additional relevant information to assist in the VCG selection process e.g., route of exposure, age, study type, test facility, etc. is extracted from the trial summary and included in the database. At the end of the data collection and curation process, all data have a SEND-like structure, either completely SEND-compliant or SEND-like, and are moved into the VICOG database.

SEND domains currently included are:
- Body weight (BW), microscopic findings (MI), and laboratory (LB) domains have been collected from most data sets.
- Clinical observations (CL), organ measurements (OM), and macroscopic findings (MA) have been gathered from only a portion of the collected data sets.
- Demographics (DM) and trial summary (TS) information.

The species included so far (as of March 2023) in the database are rat (60,000), mouse (3,000), monkey (2,500), and dog (3,500) (Fig. 2). There is a high density of data from the domains BW, LB, MI, CL, OM, and DM for the rat repository.

This is a uniquely rich database for toxicologists to understand numerous factors that are crucial to study analysis, not only within their company but also within other companies, such as background incidence of pathological findings, biomarkers, strain-specific normal values, etc. An important feature of the database is that all data points can be traced back to the individual animal.

Examples were shared illustrating the intensive process of data curation for the current VICOG database using the BW and LB domains as examples. Data curation efforts were high, as many parameters had to be further harmonized, even though these data were submitted in the standardized SEND or SEND-like format. Effect terms describing the measurements differed with regard to spelling and wording. For example, for the BW domain, there were several variations of the term “body weight” in the database, such as “BODY WEIGHT”, “BODYWEIGHT”, and “Bodyweight” not to mention the differences between body weight and terminal body weight, where the latter required a clear distinction in the database. The LB domain showed an even more heterogenous picture, e.g., for the effect “Activated Partial Thromboplastin Time”, alternative terms included “ACT.PART.THROMBOPL.TIME”, “ACT.PART.TROMBOPL.TIME” or “APTT”. The LB domain originally reported 203 unique effect terms for plasma measurements, which were reduced to 153 terms after harmonization.

The units in the database also required harmonization, as, for example, body weight was reported both in grams and kilograms (BW domain). Absolute basophil counts were reported in /µL, 10^3/µL or 10^9/L (LB domain), so units needed to be standardized. Several special characters were used for the units, such as “10^3” vs. “10E3” or “µL” vs. “/µL”, which had to be standardized. In total, 156 units were used across all data sets in the LB domain submitted to the VICOG database; these were ultimately standardized to 18 SI units.
As the database curation process is time-consuming, which is in turn costly, the most critical parameters influencing the VCG selection process should be prioritized for collection, curation, and inclusion in the database. Many parameters reported in the database may have a significant impact on the distribution of values per analyzed effect. Therefore, a partial eta squared analysis explored the impact on data variance that various parameters have on study measurements. The partial eta squared approach estimates the variance explained by a given variable after the influences of all other factors in the overall variability have been controlled (Adams and Conway, 2014).

The results for the comparisons of three parameters, namely “body weight”, “company”, and “year of study” are summarized in Table 1 using the LB domain as an example. Body weight was used as a surrogate parameter for the age of the tested animal, a parameter which is rarely reported for rat studies in the database. For instance, using the LB domain as an example, the company that submitted the data accounted for 32/30% (female/male) of the variance observed in the potassium measurements in rodents, whereas the body weight differences of the analyzed animals did not have an impact. Likewise, the year of study conduct also played a significant role in the variance of several parameters, e.g., 23/22% (female/male) for albumin measurements. On the other hand, in this dataset, differences in body weight did not have a large impact on the variance of almost any of the hematological measurements except for alkaline phosphatase (30/38%). The impact of the company and study year might be explained by differences in analytical methods and changes in experimental study conditions over time, e.g., use of different vehicles, diets, etc. Further analyses are needed, including more study parameters, to identify where hidden confounders may influence the measured parameters. This will help to define critical parameters that must be considered in the matching process to select animals for VCG generation.

### 3.2 Initial Matching Criteria and Selection Procedure for VCG Generation and Impact of VCGs on Study Decision

To identify animals within the VICOG database which are suitable to generate VCGs optimally matched with the treatment group animals, a matching pipeline must be established. The preliminary matching procedure (Fig. 3) and a first case study (Fig. 4) was presented.

The basic concept of the preliminary matching pipeline is that the study specific variables of the treatment arms of Study X are identified at the start of the study and used to determine animals from the VICOG database that are the most “similar” to the treated animals, referred to as “matched” controls. Those “matched” controls can then be used to replace totally or partially the CCG from the Study X. In the current pipeline, the study specific variables are the species, the strain, the route of administration,
Fig. 3: Preliminary Matching Pipeline with Initial Matching Criteria and Selection Procedure for VCG Generation.

Fig. 4: Distribution of AST Values in males from CCGs and VCGs in 26 studies.

The anonymized study number is displayed above each density plot. For the concurrent controls, AST distributions are shown in green, while for the pool of “matched” controls used to build VCGs the AST distributions are shown in blue. Vertical bars on each density plot are the average concentrations for each control group. The green and blue triangles displayed on the x axis are the median AST values in CCG and VCG respectively. For each study, the CCG size is displayed (e.g., for study 10, “CC=10”). The ratio of the number of matched controls identified above the total number of control animals is also displayed for each study (e.g., there were 122 matched controls identified for study 10 out of 1104 control animals in the repository used). The percentage of treated animals matching the controls is also displayed (e.g., for study 10: 100% of the animals in the 3 dose groups were matched to controls.

The year of conduct of the study, the sex, and the body weight of animals from the treatment groups at start of study. To identify the “matched” controls we use an exact matching procedure for the species, strain and sex which consists in selecting in the VICOG database the controls from the same species, strain and sex as the treated animals. For the year of the study and the body weight a coarse exact matching (CEM) algorithm was used (Iacus et al., 2008); it consists in binning the study year of the controls in bins of 5 years and the body weight in bins of 10 grams. By applying this algorithm, the year of conduct of the studies where VCGs come from is no more than 5 years past the date of the analyzed study (Study X) and the body weight of each “matched” control is equal (±5 grams) to the body weight of one or several treated animals (from Study X). If the criteria are not fulfilled, the control is discarded and not included in the VCGs suitable to replace the animals in the CCG, defined here as the pool of “matched” controls.
Fig. 5: Probability of obtaining identical statistically significant results when replacing concurrent controls with VCG, by study and sex

The red bars represent the percentage of matching statistically significant results between the concurrent controls and VCG for the low dose (Lo), while green and blue represent the corresponding percentages for the mid (Mi) and high (Hi) doses, respectively. For each study, the CCG size is displayed (e.g., for study 10, CC =10).

From study to study, the percentage of treated animals matched to VCG can differ (see Fig. 4 for examples). When there are fewer animals in the matched pool than in the CCG, no VCGs can be generated. In this case, the corresponding study cannot be reanalyzed with VCGs.

From the pool of “matched” animals, a sampling procedure draws Z number of VCGs, with each VCG containing the same number of animals as the initial control arm of the study and where Z is dependent on the number of matched controls in the pool. Each animal included in the pool of “matched” controls is included at least once in the VCGs. The smaller the pool of “matched” controls, the smaller Z will be. Typically, several hundred VCGs are drawn from the matched pool. For each VCG we compute a Shannon index reflecting the diversity of the studies from which “matched” controls are drawn (the more studies that are sampled to build a given VCG, the larger the Shannon index is for that VCG). The 100 VCGs with the highest Shannon index are then selected for reanalysis.

A case study was presented using the preliminary matching pipeline to determine whether the same significant findings (i.e statistically significant changes of the endpoints in treated compared to controls at alpha level=5%) would be reported for a study if a VCG was used in place of a CCG. The case study was built on twenty-seven legacy studies provided by Roche where rats were treated for 4 weeks with different drug candidates and changes in serum aspartate aminotransferase (AST) were measured as the endpoint. Control animal data for VCG generation was also taken from Roche legacy studies only. For one study (Study 24), insufficient “matched” male controls could be identified, as all animals exhibited a high body weight at study start (mean = 294 g versus mean = 196 g in the other studies); therefore, only 26 studies are shown in Figure 4.

In some instances, the VCG AST distributions and concurrent control AST distributions are very similar in terms of mean and standard deviation (see Study 20), while in other cases the distributions are very different (see Study 6 where the difference between AST average concentrations in CCG and VCG is ~35 U/L). For study 13, only 10 matched controls could be identified which match only 27% of the treated animals; this allowed only a single VCG to be built that was the same as the CCG size (i.e., n=10). Study 13 was run in 1995, and there were few animals with appropriate body weight from studies run in the 5 years preceding 1995 in the database, therefore only few of the treated animals could be matched. For all the other studies, at least 90% of the treated animals could be matched to controls.

Comparing parameter distributions between concurrent controls and VCGs, such as the AST distributions shown in Figure 4 is useful to understand how similar concurrent and virtual controls are, and these comparisons can also assist in identifying critical inclusion criteria for the matching pipeline.

However, the main question is whether using VCGs results in different significant findings in a study. That is, if a VCG were used in place of a concurrent control, would the overall results of the study have changed? The twenty-seven legacy study were reassessed by comparing treated arms to VCGs in place of concurrent controls (Fig. 5). In this case study, the difference in AST levels in treated compared to controls was declared different under the significance level of 5%.

In Studies 5, 13 and 17, there were fewer than 3 females with AST concentrations values reported (in the concurrent controls or treated), so female groups for those studies were not reanalyzed. In females from Study 24, results were different when
The boxplots display the power computed for each comparison for all the 27 selected studies (using either concurrent controls or VCGs), by sex and dose group (controls versus low, medium and high doses). VCG sample sizes are a multiple of the sample size of the concurrent controls (from 1, with equal sample size to 3 where VCGs are three times larger than the concurrent controls sample size).

using VCG or concurrent controls (probability = 0, in all doses groups). Only 60% of the treated females matched the controls, their body weight at study start was higher (average body weight ~ 200 g) than what is usually observed in other studies (average body weight ~ 150 g), this explains why it was difficult to match all the treated females. Furthermore, AST concentrations observed in VCGs were much lower (average concentration around 70 U/L) than in concurrent controls or treated animals (~100 U/L). Figure 4 shows that average AST concentrations in concurrent controls can vary across studies between 50 and 150 U/L, which corresponds to the reference range reported in (Sharp, 1998). One of the reasons for this, mentioned by (Everds, 2015), could be the psychological stress of the animals which could differ from study to study.

The overall probability of obtaining identical statistically significant results across the twenty-seven reanalyzed studies are the following in females: 77%, 63.2% and 51.5% for the low, mid and high dose, respectively and the following in males: 70.8%, 62.5% and 67.1% for the low, mid and high dose, respectively. For some studies, the VCGs gave the same study results as with the concurrent control, but for others, the VCGs led to a different result. The differences observed could be explained by the AST distributions which differ in terms of mean and standard deviation between the CCG and VCG. When distributions overlap well like for the Study 1 in males, the percentage of matching significant results is close to 100%. Although study results are usually based on statistical significance, the study director may override statistically significant effects if the findings are not biologically relevant, e.g. because they show no dose-dependency, are within the upper or lower limit of normal, show significant declines where only increases are of toxicological relevance, and other reasons. Therefore, it is possible that in instances where using the VCG changed the statistical results of the study, the study director would have overruled some of the results as not being treatment-related and, thus, the VCG-based results would be accurate. On the other hand, the study director may also consider effects as being treatment-related based on expertise, previous experience or based on the mode of action of the test item despite a lack of statistical significance. This is more common for non-rodent studies with smaller group sizes.

Inclusion of further parameters in the matching pipeline (e.g., vehicle, dosing regimen, etc.) may be necessary to improve VCG generation. Finally, this reanalysis was only from a single VICOG partner (i.e., Roche), so reanalysis of legacy studies by other partners must also be carried out as well as incorporation of the full VICOG database rather than just VCGs generated from a single partners’ internal database. Data analysis to address these issues is ongoing. More results on the Roche case study are available in slides presented at the Non-Clinical Statistics Conference held in Louvain-la-Neuve in 2022 (Duchateau-Nguyen, 2022).

A potential advantage of VCGs is that their sample size can exceed the concurrent control sample size; the only limitation comes from the size of the pool of available matched controls. Therefore, utilizing a VCG can increase the power of a study, which is defined as the ability to detect a difference between controls and treated groups if a difference does exist (Jones et al., 2003). A 1:1 ratio of test to VCG animals as well as increasing ratios of 1:1.5, 1:2, 1:2.5, and 1:3 were assessed to establish whether increasing the size of the VCG would increase the power of the study. As depicted in Figure 6, the median power of studies using VCGs was
slightly higher than the median power of studies using the concurrent controls even in just the 1:1 group. Not surprisingly, when increasing the size of the VCG, a trend of increased study power was observed, which is in agreement with the observation made by (Bate and Karp, 2014). In their paper, those authors show that the sensitivity of a study is maximized when the number of animals in control group is increased relative to the number of animals in the treated groups where the control group is compared to each dose group of a study.

4 Standardizing VCG Study Design Terminology

A conventional study design of a systemic toxicity study consists of a control group and three dose groups (low, medium, and high), and the original plan for VCGs was that they would replace concurrent controls outright. While this goal may still be achievable at some point, the initial proof-of-concept studies illustrate the role of confounding factors (Gurjanov et al., 2023) requiring that additional confidence must be gained in VCGs before complete replacement can be considered. As a consequence, several study designs utilizing VCGs were proposed for a stepwise approach to build confidence in the VCG process before arriving at complete replacement of concurrent controls with VCGs. These designs for prospective studies are defined below and summarized in Table 2.

4.1 Dual Control Design

The first option to implement VCGs is to retain the concurrent control and simply add the VCG as a second control; this design is termed the “Dual Control Design”. The VCG for this design would consist of the same number of animals (n) used in the concurrent control and would be generated by selecting n animals from the matched pool based on study specific parameters from the treated arms. The advantage of this approach is that it is likely the most comfortable way to introduce VCGs in toxicology studies, as the traditional backbone of a toxicology study remains but it also offers additional data in the form of the VCG. Additionally, the VCG represents a truly randomized control group of animals drawn from controls matched to the treated animals. While adding in a second control in the form of a VCG almost doubles the effort for the study director and may produce conflicting results, which require additional assessment, such a design will deliver valuable information for a qualification and the sources for conflicting results. However, if both the VCG and CCGs agreed both in absence or occurrence of changes, this would strengthen the robustness and validity of the study outcome while also building confidence in VCGs.

4.2 Hybrid Design

A Hybrid Design involves reducing the number of concurrent controls used in traditional toxicity studies and replacing them with VCGs. The intent is that the final control group will contain the conventional number of controls (e.g., 10 animals in a rat study), but some animals are concurrent controls while the remaining animals are virtual control animals (e.g., 5 concurrent and 5 virtual animals). The virtual control animals for this approach would be selected in the same manner as for the Dual Control Design: randomly selecting 5 animals from the matched pool based on study specific parameters from the treated arms. The Hybrid Design offers the advantage of continuing to accrue data for the virtual control database (allowing for detection of genetic drift or other

<table>
<thead>
<tr>
<th>VCG Study Design</th>
<th>VCG Study Design Description</th>
<th>Example of Animal Counts (provided for rodent studies)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dual Control</td>
<td>VCG is used in parallel with a traditional concurrent control</td>
<td>Treated: 10 animals/sex/dose</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Control: 10 concurrent animals/sex and a separate set of 10 virtual animals/sex</td>
</tr>
<tr>
<td>Hybrid</td>
<td>Study control group is composed of concurrent and virtual control animals to form a final study control group that is the same size as the treated groups</td>
<td>Treated: 10 animals/sex/dose</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Control: 5 concurrent animals/sex combined with 5 virtual animals/sex</td>
</tr>
<tr>
<td>Augmented A – Expanded</td>
<td>Study control group is composed of concurrent and virtual control animals to form a final study control group that is larger than the treated groups</td>
<td>Treated: 10 animals/sex/dose</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Control: 10 concurrent animals/sex combined with 10 virtual animals/sex</td>
</tr>
<tr>
<td>Augmented B – Iterative</td>
<td>Study control group is a VCG using an iterative sampling approach where Z iterations from the VCG matching pool are sampled to select a final VCG that is the same size as the treated groups</td>
<td>Treated: 10 animals/sex/dose</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Control: Z x 10 virtual animals/sex</td>
</tr>
</tbody>
</table>
time/procedure-related changes in parameter values), while allowing for the number of animals to be reduced without sacrificing statistical power.

### 4.3 Augmented Designs

The main purpose of the augmented designs is to increase statistical power of the study through data from a higher number of control animals. An augmented study design can take one of two forms. The first, referred to as Augmented A – Expanded, adds virtual control animals to conventionally-sized concurrent controls. For example, the treated arm contains 10 animals per dose while the overall control group would contain 20 animals; these 20 animals would be a combination of 10 concurrent controls and 10 virtual control animals (versus 2 groups of 10 in the dual control design). The virtual control animals for this approach would be selected in the same manner as the Dual Control and Hybrid Designs: randomly selecting n animals from the matched pool based on study specific parameters from the treated arms. Besides increasing the statistical power of the study, this approach retains the concurrent controls, allowing for the virtual control database to be updated over time. However, the drawback of this approach is that it does not reduce the number of animals used compared to a traditional toxicology study. Nevertheless, a more precise estimate of the background and smaller variability can be achieved, and for this reason, it may be easier to detect an effect.

The second approach is called Augmented B – Iterative, which uses a sampling approach that draws Z number of VCGs, with each VCG containing the same number of animals as the initial concurrent control arm of the study and where Z is dependent on the number of available matched controls (see Fig. 3). Augmented B – Iterative thus represents a full replacement of the conventional control group. Like the Augmented A – Expanded approach, this approach could also offer increased statistical power. In addition, because the various sources of variability are not completely identified and quantified, the creation of multiple (Z) VCGs is the best way to provide an estimate of the treatment effect. A drawback of this approach is that the absence of concurrent controls will prevent systematic accrual of new animal data to the database. Without alternative means of systematically updating the database, the VCGs generated from the database may lose relevance, due to unrecognized genetic drift or other temporal changes in parameter values. However, as fewer animals are used, this makes the approach attractive from both an ethical and a business perspective, as there are savings both in animal lives and study cost.

In the context of a qualification procedure of the VCG concept, most companies will apply the “Augmented B Iterative” to reanalyze legacy studies.

### 5 A Roadmap to VCG Implementation

The proof-of-concept results and subsequent breakout sessions during the workshop identified two major steps that must be taken before the VCG approach can be implemented: 1) continued database development including in-depth statistical characterization of the collected data and 2) designing and implementing a qualification procedure. These steps are highly interrelated and form a feedback loop, as information gained during database development can be used to inform the qualification procedure and vice versa. With sufficient database development and favorable results from the qualification procedure, VCGs can then be implemented in practice. The steps toward implementation are depicted in Figure 7.

First, as described further above the VICOG database needs to be developed and populated. The collected data needs to be characterized in terms of underlying confounders and drift of parameters, which might influence the optimal matching between VCGs and treatment groups. All steps need to be accompanied by quality assurance (QA) and computer system validation (CSV) to safeguard process and data integrity.

As it stands, the database development and the qualification procedure will focus on: 1) 28-day oral repeated dose studies in rats, and 2) oral dose-range finding (DRF) studies in NHPs. For rats, 28-day repeated dose oral toxicity studies are by far the most frequently performed study, so prioritizing these studies will result in sufficient data for analysis. For NHPs, DRF studies frequently do not include control groups, so the addition of VCGs will enhance these studies. Further, the NHP studies include
Baseline measurements (i.e., measurements before the start of the study, which are not available for rodents), which will be useful in the matching procedure.

Both the database development and the qualification procedure have several action items that must be addressed using both the rat and NHP before VCGs can be implemented, and these action items are outlined in more detail in the sections below.

### 5.1 Continued Development of the VICOG Database

Although the VICOG database has been under development since May 2022, additional data efforts are needed to overcome data gaps. Included in the current database are data from BW, MI, and LB domains for the majority of data sets submitted by VICOG partners. Likewise, the DM and TS are also extracted and included. For a subset of these data sets, MA, CL, and OM have also been collected.

During the workshop, other domains were identified as important, therefore, necessary for inclusion in the VICOG database, including exposure considerations (e.g., type of vehicle, frequency of dosing), vital signs, food and water consumption and ECG measurements. However, there is also some overlap with some of these domains – for example, vital signs and clinical observations have some shared parameters – so all these domains may not be needed. Other domains, like food and water consumption as well as housing conditions (e.g., single or grouped housing, bedding, etc.) were also proposed as important, but these data are not always captured in the current version of the database.

The current domains and those currently under consideration for inclusion in the VICOG database are summarized in Table 3. Consensus on which of these domains should be included will be based on analyses such as shown in Table 1.

Part of continued development of the VICOG database will also be to decide whether it will serve as the universal database or whether it will rather be used as an empty shell that will be populated by individual stakeholders with the data generated within their test facility. While the inclination for the VICOG database is that it should be a collaborative effort to define standards of data curation and statistical characterization, a possible outcome of the qualification process could provide evidence that data for VCGs can only be used if they originate from the same test facility. Indeed, the high variance observed in the hematological parameters when stratified by company (Tab. 1) suggests that company-specific databases might resolve some of the variance. Consequently, the database development exercise may show that it is not feasible to have a single, large database, but rather site-specific databases are required. However, additional analyses are necessary to draw this conclusion. The next step, therefore, is to critically assess the impact that the test facility location has on specific parameters. If the test site location is a major driver of the study result, it must be determined whether statistical controls can be put in place to account for these differences to preserve the universal database approach or if this parameter will force individual databases for each company. In any case, the database will require a continuous update with new control group data to account for time-dependent variances caused by changes in biochemical assays, changes in diet or housing, genetic drift or other aspects.

### 5.2 Qualification: Building Confidence in the VCG Framework

#### 5.2.1 Visualizing distribution and time control graphs

The first step in qualifying the VCG Framework will simply be to gain a deeper understanding of the data from the extensive VICOG database. Statistical distributions will be constructed to understand the variability of each parameter. This exercise was demonstrated in a first case study for AST (Fig. 4), and the qualification procedure will continue this work for additional parameters. Additionally, as the VICOG database includes data over several years, time control graphs will be constructed to identify parameter changes over time. These types of plots can direct additional data analysis if there is a questionable finding in the study results. Interactive web tools will be utilized to create visualizations of the distributions and temporal relationships to simplify data analysis for the end-user. It was concluded that it will neither be necessary nor feasible to construct distributions and time control plots for all assessed parameters, as many of these are interrelated and the process would be far too time-consuming. Therefore, priority will be given to such parameters that are already available in the VICOG database as well as parameters which are particularly critical for decision making (e.g., liver transaminases). Further analyses will reveal these key parameters, which shall be assessed prior to using the data for VCGs.

---

Tab. 3: Current VICOG Database Domains and Domains Under Consideration for Future Inclusion

<table>
<thead>
<tr>
<th>Collected from all data sets</th>
<th>Collected from a subset of data sets</th>
<th>Under consideration for future collection*</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Body weight (BW)</td>
<td>• Macroscopic findings (MA)</td>
<td>• Exposure (EX)</td>
</tr>
<tr>
<td>• Microscopic findings (MI)</td>
<td>• Clinical observations (CL)</td>
<td>• Vital signs (VS)</td>
</tr>
<tr>
<td>• Laboratory parameters (LB)</td>
<td>• Organ measurements (OM)</td>
<td>• Electrocardiography (ECG)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Food/water consumption</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Housing conditions</td>
</tr>
</tbody>
</table>

*These domains were identified during the workshop, but this list is not meant to be exhaustive. Additional domains may be considered, and, ultimately, the final list of domains will be decided using results from statistical characterization and industry consensus.
Hidden confounders can result in erroneous study conclusions. An example of this was discovered in the data underlying Table 1 regarding the variability of LB parameters against body weight. Two distributions per gender were observed when evaluating the variability of several hematological endpoints (i.e., reticulocyte, absolute basophil, absolute eosinophil, and absolute monocyte counts) against body weight. After confirming that this was not due to a data curation error, it was found that the source of this variability was attributable to the use of two different strains of Wistar rats – Wistar Han vs. Wistar WU. This strain distinction had not been captured in the data collection and curation process. Another example of a hidden confounder is exemplified in a case study involving clinical chemistry where calcium and potassium values exhibited a bimodal distribution (Gurjanov et al., 2023). Upon further investigation using a time-course plot, the change in electrolyte values was ultimately attributed to a change in anesthesia over time from carbon dioxide to isoflurane. These examples underscore the importance of identifying hidden confounders. Hence, the qualification procedure will include additional investigations to identify such confounders.

5.2.3 Matching criteria and guidance
Refining the matching criteria to identify the optimum virtual control animals will also be a critical component of VCG qualification, and this will begin with how to define “matched” at the individual animal level. That is, how similar is similar enough? First, the most critical parameters must be identified. As demonstrated in Table 1 and summarized in Section 3.1, body weight, which is considered of key importance for age matching (the date of birth is often not available for rodents), actually did not play a major role in the variance of most laboratory measurements in rodents in the dataset analyzed so far, but further analyses are needed. However, the year in which the study was performed did influence the variance. Additional analyses should be carried out to determine the impact each parameter has on study outcomes to refine the matching criteria for rats and NHPs. This information will also help focus database development efforts, as those high impact parameters can be prioritized for collection and curation.

Consideration must also be given to how many study parameters are needed to achieve similarity. The more parameters are compared, the lower the number of matching animals will be. A balance must be found between using sufficient parameters to identify a matching animal without exhausting the matching pool – essentially, a matching “threshold” for both rats and NHPs.

5.2.4 Reanalysis of legacy studies
Perhaps the most persuasive aspect of the VCG qualification procedure will be the retrospective analysis of legacy studies, which will compare how the study outcome is affected when using a concurrent control versus using a VCG. While the proof-of-concept re-analyzed the legacy studies based only on statistical significance, the qualification procedure will expand upon this by also evaluating concordance for biological relevance as well as overall study result concordance. The statistical results will be presented in the form of heat maps, as proposed in Figure 8.

However, the mere statistical comparison between the original data and the results obtained with VCG will not be sufficient to assess the validity of the approach. As described in Section 3.2, a study director may disregard statistically significant results because they are not biologically relevant or can identify non-significant findings as treatment-related. Therefore, the study director needs to reassess a legacy study after replacement of CCGs with VCGs and identify the changes in terms of “treatment-relatedness”. Subsequently, they need to assess whether missing or additionally detected treatment-related changes are influencing the overall conclusion of the study in terms of NO(A)EL, identified target organs, monitorability of affected parameters, and recovery from effects. The goal of the qualification procedure is to test whether using VCGs would compromise study outcome and thus endanger patient safety during First-in-Human or later clinical studies trials. Part of the qualification procedure mentioned here can be assisted by statistical analyses using contingency tables (with statistical significance (Yes or No) as rows and treatment-relatedness (Yes or No) as columns). For both CCG and VCG, such contingency tables can be constructed used to determine agreement.

Concurrent controls in legacy studies are replaced with VCGs and the statistical results are compared. Heatmaps provide a rapid overview regarding matching and missing results. Subsequently, the study director assesses which of the newly identified significances are treatment-related and which can be disregarded due to lack of toxicological relevance. In a third step, the matching and mis-matching treatment-related results are compared and assessed regarding their influence on the overall study conclusion.
5.3 Recruitment of Additional Expertise

To date, the VCG approach has been a highly collaborative effort; however, initially, participation was limited mostly to pharmaceutical partners, with expertise from study directors, toxicologists, and statisticians. While these experts have been instrumental in VCG development thus far, during the course of the workshop, it became clear that additional expertise is needed to continue to build confidence in the VCG approach. Accordingly, representatives from the following societies and subject matter experts should be included in future developments of the VCG approach. This list will likely expand as the VCG approach evolves.

- Pathologists
- Data curation and ontologies experts
- Study design experts
- Laboratory animal societies
- Good Laboratory Practice (GLP) and Computer System Validation (CSV) experts
- Machine Learning (ML)/Artificial Intelligence (AI) experts

While feedback from all of these experts is valuable, of particular importance are subject matter experts in the field of pathology. Pathologists use control slides to identify and grade treatment-related findings in traditional toxicity studies. However, the current VICOG database does not include the original control slides but is limited to the verbal description of microscopic findings, including incidence and severity (i.e., historical control data). Historical control data serve to understand the true background incidence of spontaneous lesions; as such, they are used as a complementary tool for additional perspective on findings. Consequently, it would be tremendously valuable to understand from pathology experts in which instance relying on the verbal description of microscopic data might not be sufficient and when a control slide is needed to perform comparisons, even if a validated control data repository is available. Ultimately, absence of control slides is very likely to make VCG implementation difficult, and digital slides could serve as an alternative.

Likewise, ML/AI experts may also be able to contribute significantly to the development of the VCG approach for both the database development and the matching procedures. As has been discussed, the data collection and curation process, which is the foundation of the VCG process, is laborious, which is, in turn, costly, but automatic data processing approaches may mitigate this issue. Also, after parameter distributions are better understood, “normal” ranges for parameters can be established and limits can be set to flag data points that appear to be outliers. Curation efforts can then be focused on these outliers rather than the entire database, which will drastically reduce manual efforts while still allowing the database to grow at a reasonable pace. Input from ML/AI experts will be necessary to establish pipelines for automated collection to ensure all outliers are captured, resulting in a high-quality database.

ML/AI experts could also explore unsupervised analyses to better understand how the control animals are clustering (i.e., what determines their similarity), which could be useful for prioritizing matching parameters. Additionally, these experts could advise on how to utilize ML approaches, like K-Nearest Neighbors (KNN) or random forest approaches, to evaluate the performance of similarity definitions. For example, using a set of parameters to define similarity, a “landscape” of virtual control animals would be generated where certain animals group together. Using a KNN approach, one could predict the parameters of animal X from that landscape using its K nearest animals and compare the predicted to the known parameters of that animal. The accuracy of this prediction would reflect the precision of the initial set of similarity parameters.

6 Promises of the VCG Approach

6.1 Deeper Understanding of the Data

Confident study results are built on a foundation of robust data. The creation of the VICOG database has produced such a data set, and it has great potential to help us better understand study data. For example, the large number of control animals in the VICOG database allows normal distributions for each study parameter to be developed. An example of this was demonstrated in Section 3.2 (Fig. 4). The VICOG database can also help to identify hidden confounders as described in Section 5.2.2.

6.2 Resolution of Statistical Shortcomings

When considering the appropriate group size for toxicology studies, consideration must be given to ensuring that there is a sufficient number of animals to ensure a reasonable probability of detecting adverse effects of a drug. This needs to be balanced, however, by ethical concerns that attend increasing the group size beyond what is strictly necessary. Appropriately selected VCGs have the potential to increase statistical power of the study by increasing the size of the control group above that of a typical CCG.

The proof-of-concept analysis described in Section 3.2 (see Fig. 6) demonstrates increased study power for 28-day oral studies in rodents with what is now defined as the Augmented B (Iterative) approach. Even at just the traditional 1:1 ratio of test to control animals, the power of a study would increase with VCGs generated using the iterative approach, and the study power continues to increase as the ratio of test to control animals increases. In addition, increasing the control group size using the Augmented A (Expanded) would also increase the study power. For studies that are run using non-rodent species such as NHPs, dogs, minipigs, etc., fewer animals are typically used due to prohibitive costs as well as ethical issues (Bliss-Moreau et al., 2021). Therefore, VCGs could be particularly useful for increasing study power in these instances.
Tab. 4: Potential Number of Animals Spared by Complete Replacement with VCG

<table>
<thead>
<tr>
<th>Species</th>
<th>Number of Control Animals Used Per Study (n)</th>
<th>Estimated number of drug candidates subjected to nonclinical toxicology testing annually from 2010-2019 in the US (n)</th>
<th>Total number of control animals used in nonclinical toxicology testing per year (n)</th>
</tr>
</thead>
<tbody>
<tr>
<td>DRF</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rodent</td>
<td>10</td>
<td>380*</td>
<td>3,800</td>
</tr>
<tr>
<td>GLP</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rodent</td>
<td>20</td>
<td>380*</td>
<td>7,600</td>
</tr>
<tr>
<td>Non-rodent</td>
<td>10</td>
<td></td>
<td>3,800</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td></td>
<td>15,200</td>
</tr>
</tbody>
</table>

*The FDA approved, on average, 38 new drugs per year from 2010-2019 (CBO, 2021). Considering the attrition rate of drugs from the nonclinical phase is approximately 90% (Sun et al., 2022), it can be estimated that the associated starting pool of these drug candidates at the nonclinical phase was approximately 380.

6.3 Reduced Animal Use

Undoubtedly, the most obvious advantage of employing VCGs in nonclinical testing is the substantial reduction in animal use. While the exact number of animals used in scientific research each year in the US is unknown, as reporting the number of mice and rats used in laboratory experiments is not required under the Animal Welfare Act (AWA), the number of animals used in nonclinical testing can be estimated using data on drugs approved by the FDA.

According to the Congressional Budget Office, on average, the FDA approved 38 new drugs per year from 2010 to 2019 (CBO, 2021). However, it is estimated that greater than 90% of drugs fail to progress past the nonclinical phase of the drug discovery process (Sun et al., 2022). Based on this attrition rate, it can be estimated that these 38 drugs came from a pool of at least 380 drug candidates per year, all of which would have generally been subjected to nonclinical toxicology testing. During the nonclinical phase of drug development, at minimum, each drug candidate will typically undergo a dose-range finding (DRF) study in at least two species (one rodent and one non-rodent) and a 28-day repeated dose toxicity GLP study in two species (one rodent and one non-rodent). There is currently no standard approach for DRF studies; however, assuming even 5 rodents per sex are used as control animals for DRF studies, utilizing VCGs could result in a savings of up to 3,800 animals per year (see Tab. 4). For GLP studies, 10 rodents per sex per dose are used while a smaller number of animals are used for non-rodent (usually NHP) studies, again due to ethical and cost restrictions. Using these estimates and assumptions, incorporating VCGs into GLP studies can potentially spare up to 7,600 rodent and 3,800 non-rodent (NHP) lives on average per year for drugs submitted to the FDA.

It is important to emphasize that these estimates are only for drugs submitted to the FDA and relate only to the toxicity studies before first-in-human trials. If the VCG approach is adopted globally and extended to further study types (e.g., embryo-fetal development studies), the number of animals that can potentially be saved can be even higher.

Not only is implementing VCGs into nonclinical safety an ethical choice, but it is also a wise business decision. Reducing the number of animals used in a study may contribute to reducing the costs associated with testing. Assuming that roughly one-fourth of the FDA drug approvals represent biologics (Makurvet, 2021), 25% of the non-rodents calculated in Table 4 can be assumed to be NHPs. Given the current costs of up to $50,000 USD each, the cost savings for 950 NHPs would add up to $47,500,000 USD per year.

6.4 Industry Support and Regulatory Interest

Although still an emerging and evolving concept, consideration of VCGs, in some form, by regulatory authorities in the relatively near future appears to be promising. This was concluded from the results of a poll taken at the beginning and end of the workshop, where attendees were queried as to whether they thought VCGs would be accepted by regulatory agencies and in what timeframe.

The results revealed that, even at the beginning of the workshop, all respondents believed that VCGs had a chance to be implemented and speculated that regulatory agencies would accept the concept within 3-5 years. The optimistic nature was sustained through the end of the workshop, where during the follow-up poll, again, almost all respondents agreed that the VCG concept has a chance to be implemented, and the timeframe to acceptance by agencies was again predicted to be 3-5 years; however, there was an increased perception at the end of the workshop that implementation was likely to impact primarily non-GLP studies (Fig. 9).

This theme has been observed at other workshops as well. At a 2021 workshop with members of the FDA and industry that focused on how to reduce reliance on NHPs, it was noted that, for GLP studies using NHPs, eliminating CCGs and replacing them with control groups generated from historical control data – essentially, VCGs – could be a way to reduce reliance on NHPs (Ackley et al., 2023).

Although in their infancy, virtual controls have already been explored as alternatives to the standard of care in randomized control trials (RCT), as summarized nicely by (Strayhorn, 2021). Many of the advantages to be gained in nonclinical safety by VCG have already been highlighted as advantages of virtual controls in the context of RCT, including identification of accurate predictors of the outcome and achieving results more quickly (Strayhorn, 2021). The advantage of faster results can be especially useful for nonclinical safety studies relying on NHPs; this species has recently been difficult to obtain, which results in tremendous delays for studies requiring NHPs.
7 Challenges for VCG Implementation

7.1 Data Efforts
Though the VICOG database is currently quite useful due to the large amount of data it already contains, the reality is that continuing to build and maintain such a large, high-quality database is going to be extremely labor intensive. There is a perception that because SEND exists, nonclinical data are harmonized and, therefore, data collection and curation for VCG implementation should be relatively simple and straightforward. However, as demonstrated in Section 3.1/Table 1, variation in data even in standardized formats, like SEND, still exists. For example, while data are well-harmonized under the larger umbrella terms outlined by SEND such as species, strain, body weight, etc., details like units or body weight vs. terminal body weight are not always consistent. In fact, by the end of the workshop, data collection and curation efforts were perceived by the attendees as the largest hurdle to VCG implementation. Data harmonization has been identified a key hurdle for mining big data by others (Carfagna et al., 2020; Clark and Steger-Hartmann, 2018).

Also, as some of the studies that feed into the VICOG database are legacy studies conducted pre-SEND, the availability of the data in LIM (laboratory integrated management) systems may be limited, making it necessary to go back to the original reports to obtain certain information, which is extremely cumbersome. Consequently, the data collection and curation process for VCGs is and will continue to be labor-intensive and time-consuming, which, in turn, will be costly. However, it is possible that, in the future, the process can potentially be automated, reducing some of these efforts.

The data curation process may also create hurdles for uptake of the VCG approach in GLP studies down the road. Although the VCG approach will initially be implemented in non-GLP studies, ultimately, the goal is to incorporate VCGs into GLP studies. However, under GLP, any changes made to raw data during the data curation process would require a stringent and comprehensive validation procedure to assure data integrity. Consequently, very thorough records would be necessary to maintain GLP compliance, including computer system validation for all involved processes, i.e., database, statistical characterization and matching procedure and visualization tools. Maintaining this level of detail would be an arduous task; however, if VCGs improve nonclinical efficiency, the tradeoff may indeed be worth the effort.

Additionally, if it is determined that use of VCGs will require the availability of high-quality digital scans of the histopathology slides for pathologist review, this will greatly increase the time and expense of this effort, due to the need to appropriately scan and upload the slides. There will also be increased costs associated with data storage requirements as well as infrastructure for transferring and viewing the terabytes of data that a typical toxicology study would generate (Ying and Monticello, 2006).

7.2 Matching Criteria and Selection Procedures
Establishing the matching criteria is going to be an incredibly demanding exercise, as a considerable number of parameters will need to be explored during the qualification procedure to determine which are most impactful. Furthermore, if many parameters are considered high impact, this will increase the amount of data needed, but the VCG approach is not meant to be an ever-expanding effort of data collection and curation. There may need to be a tradeoff between what is scientifically ideal and what is operationally practical.

Another concern is hidden confounders. The proof-of-concept presented during the workshop and recent related publications (Gurjanov et al., 2023) has demonstrated how certain parameters that impact study outcomes may not have been reported nor collected nor even obvious that they will impact the outcome of the study (e.g., change in anesthetic procedures over time, which in turn influences determination of electrolyte levels). Therefore, despite best efforts to match virtual animals with treated groups, hidden confounders can prevent optimum matching, which will be a significant challenge for VCG implementation.
The Augmented-B approach addresses this challenge by creating multiple VCGs. However, the additional investigations into parameter impact during the qualification procedure will hopefully minimize the risk for this.

The selection procedure will also be challenging in light of GLP requirements, particularly for the augmented iterative approach. Although it offers increased power, it is a statistically complex method. Consequently, study directors would be tasked with sifting through large amounts of statistical information. Though this process could potentially be automated in the future, this would initially increase the amount of time it takes to review a study, which always translates to increased cost. Furthermore, assessing study outcomes based on multiple distributions may require study directors to make difficult decisions if some of the distributions result in statistically significant findings while others do not. For example, if for the VCG there are 10 distributions for AST and 1 of the 10 distributions produces a significant finding, is this cause for concern? What about 2 out of 10? Obviously, exploring these issues will be part of the qualification procedure, but the reality is that not all of these questions can be answered with data science alone and will require expert judgement, and, in the end, this will be a tremendous responsibility for the study directors. A final concern with the iterative approach is that it could potentially prevent traceback to the individual animal or at least increase the efforts required to traceback to the individual animal, and traceability will be necessary if VCGs are to be implemented in GLP studies in the future. While this could be overcome by detailed record-keeping, it would increase the level of effort substantially.

7.3 Governance

With the establishment of a large, collaborative database comes the question of how such a database will be governed, and this can potentially be a major obstacle on the path to VCGs because it requires many stakeholders with competing interests to agree to the bylaws. Although certainly not exhaustive, some of the questions to be answered that are related to governance are:

- Who will govern the database?
- Who will update the database?
- How will data quality be maintained?
- How long will partners have to commit to contributing to the database?
- Who has access to the database?
- Where will the database be hosted?

Due to the conflict of interest with a pharma company or a CRO, the database will likely have to be governed by a public partner who will work as a broker on behalf of all stakeholders, which has already been successfully implemented in the two IMI projects mentioned in the introduction (Briggs et al., 2012; Pognan et al., 2021). Even then, several questions remain. First, who will update the database? It will need to be decided whether raw data will be shared with the broker, who will then update the database, or if the individual pharma partners or CROs will be required to enter new data themselves. Data quality must also be maintained, and, again, the party responsible for this task (i.e., the broker or the pharma partner/CRO) must be resolved. And, of course, all of these issues are dependent on whether a stakeholder maintains their commitment to the database.

Part of the governance issue will also be accessibility. The data will be contributed by pharma and CROs, but who will have access to the data? The obvious thought is to make this an incentive driven approach, with only those partners who donate data allowed access to the database. On the other hand, to get universal acceptance of VCGs, as many pharma partners and CROs as possible will need to be on board, and enticing them with access to the VICOG database, whether they contribute to it or not, may actually increase participation. A related question is whether the database will, in whole or in part, be accessible by the public? The database could be an incredible tool for private-public partnerships, but all stakeholders would have to agree to data sharing practices.

While all these questions can certainly be answered, the answers may not be unanimous, and it will take quite some time to resolve all the issues. A possible outcome could be that instead of a central database collecting data from numerous test facilities, an empty shell of a database will be provided, which is then populated by the individual test facilities. It is for these reasons that database governance will be a considerable challenge for VCGs to overcome.

7.4 Issues that Cannot be Addressed Without Concurrent Controls

Over time, control animals undergo genetic drift, and this can be a major challenge to implementing VCGs. Without the use of concurrent controls, these important shifts over time may not be detected. Consequently, changes that occur as a result of genetic drift could be falsely attributed to the drug candidate, thus detrimentally affecting the outcome of nonclinical studies. VCGs simply cannot address these issues, so it may be the case that at least some form of a concurrent control will always be necessary to serve as sentinel animals.

Another important consideration when concurrent controls are replaced by VCGs is that there are studies during which something that was not captured in the planned study design occurred and impacted the study results for all animals including controls. In these cases, CCG data may differ significantly from the VCG and historical control data; however, the CCG data may be more relevant to interpretation of the concurrent study as it reflects the impact of the unplanned aberration that occurred during the study, likely affecting all treatment groups. Complete replacement of concurrent controls with VCGs creates a risk that study-specific effects will be interpreted as treatment-related effects.

It is also important to remember that the VICOG database is a collection of historical control data, which were, at some point, concurrent control data. So, an obstacle to at least complete replacement of VCGs is the fact that concurrent controls feed the database. As a result, if concurrent controls are entirely eliminated, new data would not be added to the database. While this could potentially be remedied by using a full set of concurrent controls that are run less frequently (e.g., running a CCG parallel to
Tab. 5: Workshop Action Points

<table>
<thead>
<tr>
<th>Action Item</th>
<th>Action Plan</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Database Development</strong></td>
<td></td>
</tr>
<tr>
<td>Establish which domains/parameters will be included in VICOG database</td>
<td>• Data analysis of current VICOG database to identify high priority domains/parameters&lt;br&gt;• Establish industry consensus for high priority domains/parameters using expert judgement</td>
</tr>
<tr>
<td>Resolve whether a single VICOG database or company/site-specific VCG databases will be constructed</td>
<td>• Determine the magnitude with which a company/site impacts domain/parameter variability&lt;br&gt;• If the company/site has a large impact on variability, determine whether statistical controls can be put in place to account for differences</td>
</tr>
<tr>
<td><strong>Qualification Procedure</strong></td>
<td></td>
</tr>
<tr>
<td>Characterize parameter distributions</td>
<td>• Construct distributions for critical parameters</td>
</tr>
<tr>
<td>Characterize temporal relationships</td>
<td>• Construct time control curves for critical parameters</td>
</tr>
<tr>
<td>Identify hidden confounders</td>
<td>• Identify parameters with high variability and prioritize for further investigation&lt;br&gt;• Assess time control curves for changes in parameters over time</td>
</tr>
<tr>
<td>Refine matching criteria</td>
<td>• Characterize the impact of study parameters&lt;br&gt;• Evaluate the influence of the number and type (i.e., high, medium, low impact) of parameters</td>
</tr>
<tr>
<td>Refine selection procedure</td>
<td>• Evaluate proposed approaches (e.g., top n, top n/2, top 2n, Z iterations, etc.) for each VCG study design to determine the final number of animals used in each VCG study design</td>
</tr>
<tr>
<td>Reanalyze legacy studies</td>
<td>• Provide comparisons for VCG vs. CCG for each study design with regard to&lt;br&gt;  o Statistical relevance&lt;br&gt;  o Biological relevance&lt;br&gt;  o Overall study conclusions</td>
</tr>
</tbody>
</table>

*All action items must be completed for both 28-day oral repeated dose rat studies and DRF NHP studies.*

---

Fig. 10: Workflow for Education and Policy Change

---

As with any novel technology, educating end-users about the VCG concept and how to properly implement it will be essential to its uptake and long-term success (Fig. 10). This will first entail simply spreading the word about the potential use of VCGs in nonclinical safety, as end-users need to become familiar with the concept before it can be successfully implemented. This will likely involve presentations at society meetings or webinars to educate stakeholders about VCGs. Furthermore, the results of the qualification procedure will need to be shared – either through oral presentations or peer-reviewed articles – to illustrate the potential and, thus, increase confidence in the VCG approach.

---

treatment groups of different studies), the reality is that the VICOG database relies on concurrent controls, and, without these, it will be impossible to maintain the database.

8 Education and Policy Change

As with any novel technology, educating end-users about the VCG concept and how to properly implement it will be essential to its uptake and long-term success (Fig. 10). This will first entail simply spreading the word about the potential use of VCGs in nonclinical safety, as end-users need to become familiar with the concept before it can be successfully implemented. This will likely involve presentations at society meetings or webinars to educate stakeholders about VCGs. Furthermore, the results of the qualification procedure will need to be shared – either through oral presentations or peer-reviewed articles – to illustrate the potential and, thus, increase confidence in the VCG approach.
Following harmonization of the approach, VCG implementation will begin in-house – that is, pharmaceutical companies and CROs will start applying VCGs internally to gain experience before including them in formal reports submitted to regulatory agencies, which will require training of study directors and close interaction with data scientists. To familiarize regulators with VCGs in practice, investigational new drug (IND) submissions should include VCGs using a Dual Control Study Design for DRFs in their study report submissions. This arrangement is the lowest risk to new drug sponsors and the most straightforward approach for regulators, as it allows them to gain familiarity with VCGs while also retaining the traditional concurrent control. Assuming successful resolution of the scientific and procedural hurdles noted above, the focus would shift to providing guidance to stakeholders concerning when and how VCGs can be used in nonclinical studies conducted to support clinical testing and marketing of pharmaceuticals. Ultimately, it would be desirable to have internationally harmonized guidance, such as that developed under the auspices of the International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH).

9 Conclusions

VCGs hold much promise for nonclinical safety studies, including a better understanding of nonclinical data, increased study power, and reduced animal use. This makes VCGs an attractive alternative to concurrent controls. However, before they can be fully implemented, there is work to be done to increase confidence in their utility – in particular, fully validating that they can be used in a manner that does not compromise study interpretation. It became evident during the workshop that overcoming the identified hurdles and challenges requires a joint effort of all stakeholders, ideally in a consortium approach.

The workshop identified areas for continued research and a roadmap for the path forward, namely 1) continuing to build the VICOG database and 2) qualifying the various VCG approaches by reanalysis of the legacy studies. The inventory of action items from the workshop is outlined in Table 5.

All of these additional efforts will be implemented for 28-day oral rat repeated dose studies, and oral DRF studies for NHPs. Although it will likely take a couple of years to expand the database and complete the qualification procedure, it is mandatory to characterize the performance of the VCG approach against concurrent controls to truly understand whether VCGs can serve as a suitable replacement for concurrent controls.

Even though there is work to be done and the implementation of VCGs will require a new way of thinking, the consensus of the workshop was positive, and there was support for the idea of incorporating VCGs into nonclinical studies in the relatively near future.

Conflict of interest

The authors declare no conflict of interest.

Funding sources

This workshop was funded by the transatlantic think tank for toxicology (t4), a collaboration of the toxicologically oriented chairs in Baltimore, Konstanz, and Utrecht sponsored by the Doerenkamp-Zbinden Foundation.

Acknowledgements

The conference organizing committee and attendees would like to thank Camila Sgrignoli Januario and Anywn Statnick for their assistance in organizing the workshop. In addition, the statistical support provided by Paolo Piraino and Matteo Piraino with Data Lab for Research and Innovation, Organon SRL, Bucharest, Romania and Rupert Kellner with Fraunhofer ITEM are acknowledged.

Author contributions

TS-H and TH organized the workshop. EG drafted the manuscript. TS-H, FB, SE, and GD-N contributed data analyses and figures. All other authors contributed to the final version of the manuscript. All attendees participated in the workshop, either in-person or virtually.

Disclaimer

The views expressed in this paper are those of the authors and do not necessarily reflect those of the Food and Drug Administration. All procedures performed on animals were in accordance with regulations and established guidelines and were reviewed and approved by an Institutional Animal Care and Use Committee or through an ethical review process.
References
Pogna, F., Pogna, F., Steger-Hartmann, T. et al. (2023). Intronsafe: Introducing the concept of virtual control groups preclinical toxicology testing. ALTEX 37, 343-349. doi:10.14573/altex.2001311
Pogna, F., Pogna, F., Steger-Hartmann, T. et al. (2020). Introducing the concept of virtual control groups preclinical toxicology testing. ALTEX 37, 343-349. doi:10.14573/altex.2001311
Steger-Hartmann, T. and Clark, M. (2023). Can historical control group data be used to replace concurrent controls in animal studies? Toxicol Pathol 0, 01926233231208987. doi:10.1177/01926233231208987

ALTEX, accepted manuscript
published December 1, 2023, version 3
doi:10.14573/altex.2310041