Quantifying the Reproducibility and Variability of In Vivo Guideline Toxicology Studies

A. Kreutz[1] A.L. Karmaus[1], O.B. Oyetade[1], K. Paul Friedman[2], D.G. Allen[3], E.N. Reinke[1], M. Paparella[4], N.C. Kleinstreuer[5]

[1]Inotiv, RTP, NC, United States; [2]US EPA, Office of Research & Development, RTP, NC, United States; [3]ICCS, Raleigh, NC, United States; [4]Medical University Innsbruck, Innsbruck, Austria; [5]NIH/NIEHS/DTT/NICEATM, RTP, NC, United States

Background and Purpose

Application of in vitro assays to toxicological hazard assessment will require contextualization of these assays relative to robust and reliable reference data. Guideline in vivo toxicology studies have historically been used for chemical safety assessments for regulatory decision-making and thus are typically the standard against which new approach methodologies (NAMs) are evaluated. Guidelines have been in practice to help ensure results of these studies can be replicated, easily interpreted, and used for actionable outcomes. However, retrospective analyses have revealed substantive variability within and among these studies. This variability, which has many potential sources, can confound the use of data from in vivo guideline studies to establish confidence in NAMs. We have compiled results from published retrospective analyses of in vivo toxicological studies to characterize quantitative variability and qualitative (i.e., hazard classification) reproducibility across several guideline study types.

Methods

First, a literature survey was conducted by querying PubMed (including MEDLINE) and Causaly databases using medical subject headings (MeSH) terms and keywords including "variability", "reproducibility", and their variants, combined with "in vivo", "animal studies", "experimental studies" and other relevant synonyms to identify existing reports that characterized both quantitative and qualitative variability in in vivo studies. All resulting literature was reviewed. To streamline our assessment, publications summarizing reviews of established toxicological test guideline studies were prioritized. Variability was defined as quantitative variance (i.e., of study outcomes, namely points of departure such as LO(A)EL, LD50) and reproducibility as qualitative concordance in hazard classification. Examples of test methods for which relevant evaluations were identified included repeated dose toxicity studies (subacute, subchronic, chronic), uterotrophic and Hershberger assays for endocrine activity, rat acute oral lethality studies, and acute skin and eye irritation studies. Each retrospective analysis retrieved from the literature was conducted independently, and our efforts to compare and summarize across findings included delving into the approaches applied for variability and reproducibility assessment, as each was optimized to be fit for purpose per dataset analyzed. Though there are different metrics reported and different approaches applied, we have developed a harmonized qualitative endpoint/hazard classification schema to make direct comparisons where possible.

Results

Retrospective evaluations identified in our literature search consistently noted that most toxicological studies have not been routinely characterized for variability metrics and emphasized the challenge with reproducibility of hazard classification. For example, hazard

classification based on data from an acute toxicity study (whether oral lethality, skin, or eye irritation) often had less than a 50% likelihood to yield the same classification as a previous study, particularly when the original test characterized the substance as having a mild to moderate effect. By reviewing data and aggregating variability and/or reproducibility metrics where available, we provide a centralized viewpoint offering a quantitative perspective on the robustness of hazard categorization resulting from these tests. In this example, acute testing studies consistently demonstrate weak reproducibility in moderate effects across numerous endpoints suggesting hazard classification of moderate acute effects may be unreliable, and NAMs may be most effective in delineating toxic vs. non-toxic binary classification.

Quantitative variability measures, where available, were also aggregated, with work ongoing to try and standardize databases to yield harmonized variability metrics that are directly comparable between study types. Ultimately, these standardized variability measures could be integrated into assay characterization to establish more realistic metrics to benchmark NAM performance.

Conclusions

Quantitative variability and categorical reproducibility are important considerations to determine if a NAM is as good or better than the existing in vivo test method. An improved characterization of in vivo variability and reproducibility metrics will support realistic expectations when building confidence for the use and interpretation of NAMs. The level of concordance between a NAM and an in vivo test is inherently limited by the extent to which the in vivo test can reproduce itself. Recognition of this reality will shift expectations away from exact concordance between a NAM and an in vivo test assessing the same endpoint. Quantifying variability is essential context to aid in understanding study performance and should be integrated into data reporting for NAMs to build confidence for NAM adoption in regulatory use. *This project was funded in whole or in part with federal funds from the NIEHS, NIH under Contract No. HHSN273201500010C and with financial support from the Austrian Federal Ministry of Environment, Department V/5—Chemicals Policy and Biocides. The views expressed in this abstract are those of the authors and do not necessarily represent the views or policies of US EPA.*