Published in final edited form as:

Reprod Toxicol. 2019 October ; 89: 145–158. doi:10.1016/j.reprotox.2019.07.012.

ToxRefDB version 2.0: Improved utility for predictive and retrospective toxicology analyses

Sean Watford[*,†], **Ly Ly Pham**[‡,†], **Jessica Wignall**[§], **Robert Shin**[§], **Matthew T. Martin**[¶,†], **Katie Paul Friedman**[†]

[*]ORAU, contractor to U.S. Environmental Protection Agency through the National Student Services Contract

[†]National Center for Computational Toxicology, Office of Research and Development, US Environmental Protection Agency

[‡]ORISE Postdoctoral Research Participant

[§]ICF, Burlington, VT

[¶]Currently at Drug Safety Research and Development, Global Investigative Toxicology, Pfizer, Groton, CT

Abstract

The Toxicity Reference Database (ToxRefDB) structures information from over 5,000 *in vivo* toxicity studies, conducted largely to guidelines or specifications from the US Environmental Protection Agency and the National Toxicology Program, into a public resource for training and validation of predictive models. Herein, ToxRefDB version 2.0 (ToxRefDBv2) development is described. Endpoints were annotated (e.g. required, not required) according to guidelines for subacute, subchronic, chronic, developmental, and multigenerational reproductive designs, distinguishing negative responses from untested. Quantitative data were extracted, and dose-response modeling results for nearly 28,000 datasets from 400 endpoints using Benchmark Dose (BMD) Modeling Software were generated and stored. Implementation of controlled vocabulary improved data quality; standardization to guideline requirements and cross-referencing with United Medical Language System (UMLS) connects ToxRefDBv2 observations to vocabularies linked to UMLS, including PubMed medical subject headings. ToxRefDBv2 allows for increased connections to other resources and has greatly enhanced quantitative and qualitative utility for predictive toxicology.

Keywords

in vivo toxicology; predictive toxicology; toxicology database

Corresponding author: Katie Paul Friedman, 109 T.W. Alexander Drive, Mail Drop D143-02, Research Triangle Park, NC 27711, paul-friedman.katie@epa.gov, Tel. 919-541-0660, Fax. 919-541-1194.

Publisher's Disclaimer: Disclaimer: The United States Environmental Protection Agency (U.S. EPA) through its Office of Research and Development has subjected this article to Agency administrative review and approved it for publication. Mention of trade names or commercial products does not constitute endorsement for use. The views expressed in this article are those of the authors and do not necessarily represent the views or policies of the US EPA.

1 Introduction

With an increasing need for rapid screening and prioritization of chemicals for hazard and risk evaluations, researchers are developing new strategies for predicting chemical toxicity. In accordance with these efforts, the Toxicity Forecaster (ToxCast) research program [1] has been developed by the U.S. Environmental Protection Agency (EPA) to assist in the realization of the National Research Council's (NRC) vision for improving the toxicity testing of chemicals for human health [2, 3]. These efforts served as an impetus to develop the Toxicity Reference Database (ToxRefDB), a digital resource of *in vivo* toxicity study results. ToxRefDB comprises information from over fifty years of *in vivo* toxicity data. The database includes information for over 1,000 chemicals, and is being used as a primary source of data for evaluating efforts of the ToxCast program [4, 5], as well as for numerous predictive and retrospective analyses [6–9]. The utility of ToxRefDB to predictive toxicology is clear; it has been used as the basis for evaluating new approach methods (NAMs) to identify specific adverse outcomes of interest [6, 10–12], as a retrospective benchmark for predictive performance of alternative approaches [7, 9, 13, 14], and in evaluation of the reproducibility and interpretation of observed *in vivo* outcomes [15, 16]. ToxRefDB has been used for a wide variety of applications across industry, government, and academia, with 41 other publications citing either Martin 2009a or Martin 2009b in PubMed as of October 2018. Thus, ToxRefDB represents a seminal resource for predictive toxicology applications, and lessons learned from the initial implementation have been addressed in a major re-development that we describe herein as ToxRefDB version 2.

To understand this re-development, it is necessary to describe previous development and the evolution of ToxRefDB. The first version of ToxRefDB (ToxRefDB v1) initially captured basic study design, dosing, qualitative information for effects, and point of departures (PODs) from summaries of roughly 400 chemicals tested in over 4,000 registrant-submitted toxicity studies, known as data evaluation records (DERs), from the U.S. EPA's Office of Pesticide Programs (OPP). These studies adhered to Office of Chemical Safety and Pollution Prevention (OCSPP) 870 series Health Effects testing guidelines. As this resource was intended to serve as training information in understanding the utility of NAMs like those employed in ToxCast [22, 23], the chemical selection for ToxRefDB was originally prioritized to maximize the overlap with ToxCast phase 1 chemicals (ToxCast ph1v1) [24], which were compiled based on commercial availability, solubility in dimethyl sulfoxide, chemical structural features suggesting diversity, and the availability of *in vivo* data, with the result that pesticide active ingredients comprised a high percentage of the ToxRefDB and ToxCast ph1v1 libraries. Expanded efforts in data collection and curation, driven by an attempt to cover as much of the primary ToxCast chemical library as possible, increased the chemical and biological coverage of ToxRefDB v1.3 to over 5,900 i*n vivo* toxicity studies from additional sources, including the National Toxicology Program (NTP), peer-reviewed primary research articles, and pharmaceutical preclinical toxicity studies, among others, for a total of over 1,000 chemicals. As an update to ToxRefDB v1, ToxRefDB v1.3 was released in 2014 to the public as three spreadsheets that consolidated information on adverse effects from the database as well as study citations [25, 26].

Though ToxRefDB is unique in its public availability, level of curation, and coverage of chemicals and study types, since the initial release of ToxRefDB v1 in 2009 and through subsequent updates, challenges have surfaced surrounding the extraction, storage, and maintenance of heterogeneous *in vivo* toxicity information. Several stakeholders commented on challenges in using ToxRefDB with respect to the vocabulary used to describe effects, including concerns about grouping effects as "neoplastic" or "non-neoplastic," [27], as well as the need to be able to integrate the data from ToxRefDB with other public databases [28]. Further suggestions about the need for negatives, i.e. chemicals tested and shown to be negative for a specific endpoint or effect, to form balanced datasets for predictive modeling [6, 11, 15, 29] strongly relate to the need for an updated vocabulary and determination of which effects are measured in a given study. Apart from better endpoint annotations, suggestions were made to update the quantitative dose-response information to allow benchmark dose modeling [30] and provide POD estimates less dependent on specific dose selection. Developmental and reproductive effects, involving complex study designs with multiple generations, also appeared to require a more complex database structure to distinguish effect levels between generations [10]. A more nebulous problem that is common to all databases that seek to make legacy information computationally accessible is minimizing data entry error rate. While error rates never reach zero, data quality could be improved through standardized form-based data extraction as part of a QA process [31].

In this work, we describe the development of ToxRefDB v2, including a detailed description of the new content and enhancements to the database. The goal of ToxRefDB v2 is to provide a public database that better supports the needs of predictive toxicology by increasing the qualitative and quantitative information available and by facilitating the interoperability of legacy *in vivo* hazard information with other tools and databases. Note that ToxRefDB v2 contains the same studies as before, but with major additions such as quantitative dose-response data. Recognizing that predictive toxicology will require iterative efforts to build computational resources like ToxRefDB, work to generate ToxRefDB v2 has been conducted primarily in four main areas:

- Aggregation of complex and heterogeneous study designs;
- Controlled vocabulary for accurate data extraction, aggregation, and integration; and,
- Quantitative data extraction, including treatment group size, incidence and effect data, and error information (e.g., standard deviation, standard error) where provided; and,
- Efforts to improve data quality (including quality assurance, QA, and quality control, QC).

This work represents a significant advancement in increasing the richness of information available for predictive and retrospective analyses from ToxRefDB.

2 Methods and Results

2.1 ToxRefDB Overview

Like ToxRefDB v1, ToxRefDB v2 contains summary information for over 5,900 studies labeled “acceptable” for data extraction purposes only, i.e. source document was readable and study design was clear, from five main subsources: DERs from the US EPA OPP (OPP DER), a subset of available NTP study reports (NTP), the open literature (OpenLit), donated pharmaceutical industry studies (pharma), and other (Other; including unpublished submissions and unknown sources) (Figure 1A). The study types included in ToxRefDB v2 cover the same study designs as ToxRefDB v1: chronic (CHR; 1-2 year exposures depending on species and study design) bioassays conducted predominantly in rats, mice, and dogs; subchronic (SUB; 90 day exposures) bioassays conducted predominantly in rats, mice, and dogs; subacute (SAC; 14-28 day exposures depending on the source and guideline) bioassays conducted predominantly in rats, mice, and dogs; developmental bioassays (DEV) conducted predominantly in rats and rabbits; multigeneration reproductive toxicity studies (MGR) conducted predominantly in rats; reproductive (REP) toxicity studies conducted largely in rats; developmental neurotoxicity (DNT) studies conducted predominantly in rats; and a small number of studies with designs characterized as acute (ACU), neurological (NEU), or “other” (OTH) (Figure 1B). Though ToxRefDB v2 contains summary data from roughly the same number of studies and chemicals as ToxRefDB v1, substantial additions that increase the utility and quality of these data have been made.

2.2 Aggregation of complex and heterogeneous study designs

Animal studies are designed to address specific hypotheses, with flexibility in study design required to potentially reduce cost, time, and the number of animals needed [32]. However, this flexibility presents challenges in structuring the study-related information, so when designing a database to capture both study design and adverse effect-related information, a structure that allows for that flexibility is necessary. An example of the needed flexibility in terms of archiving information on many treatment groups becomes apparent in consideration of the MGR study in rats [33] (Figure 2). The addition of quantitative data required a reorganization and expansion of the previous database. Thus, the number of tables, and their connections, has significantly increased in ToxRefDB v2 to enable archiving of information from heterogeneous study designs that also may have additions or deletions of treatment groups or doses needed to thoroughly investigate toxicity and complete a study (Figure 3). Figure 3 illustrates how the database schema has been modified and tables (or group of tables) that have been added with the development of ToxRefDB v2.

2.3 Controlled effect vocabulary for accurate data extraction, aggregation, and integration

2.3.1 Controlled effect vocabulary—A controlled effect vocabulary is critical for any resource aggregating information across a diverse set of sources for efficient retrieval and to enforce semantics, especially within biology [34]. ToxRefDB exemplifies this need as it is used for modeling efforts and retrospective analysis. One of the most significant challenges in extracting and/or integrating *in vivo* toxicity studies is the lack of adherence to controlled vocabularies. Inconsistencies in vocabulary arise both as advancements are made to better understand adverse effects in the fields of pathology and toxicology and through preferential terminology in reporting due to differences among experts [35]. These inconsistencies can also be seen across studies adhering to the same guideline but conducted years apart. Without adherence to a standard vocabulary that is actively updated and maintained, these studies can only be manually integrated in a way that is unreliably subjective [36]. Current interest in the use of controlled pathology terminology appears to be growing, with ongoing development of more terminologies for specific species and lesion types available via the Society of Toxicologic Pathology (STP) as part of the International Harmonization of Nomenclature and Diagnostic Criteria (INHAND) project [37]. The medical science field has made progress in the use of controlled terminology via adoption of electronic medical records and electronic health records for reporting adverse events, including data reporting from clinical trials for pharmaceuticals and medical devices [38, 39]. In fact, current efforts are underway to develop international standards for capturing data from clinical trials, which includes non-clinical data. These efforts are led by the Clinical Data Interchange Standards Consortium (CDISC), where collaboration between international regulatory agencies and their stakeholders is fostered to develop standards for digital submission of clinical trial data, with ongoing releases of improvements to available controlled terminology [40, 41].

Originally, ToxRefDB vocabulary for endpoints distinguished between non-neoplastic and neoplastic lesions, which conformed to the vocabulary used by NTP [42]. Improving the controlled endpoint vocabulary for ToxRefDB was a particular challenge because the terminology found in OCSPP guidelines or NTP study specifications may not necessarily match the reported pathology, clinical chemistry, and toxicology study results, where terminology is sometimes more specific. Guideline language needs to be flexible and lasting, rather than overly prescriptive, but this needed flexibility also leads to potential mismatching of information across studies. One demonstrative example is provided by the terminology of the guideline requirement for OCSPP 870.4100, “full histopathology on the organs and tissues…of all rodents and nonrodents in the control and high-dose groups, and all rodents and nonrodents that died or were killed during the study” [43], which doesn’t distinguish between non-neoplastic and neoplastic lesion types nor detail all possible histological findings that could be observed, e.g., hypertrophy, adenoma, fatty changes. In ToxRefDB v1, the effect vocabulary was generally standardized and hierarchically structured into broader categories called endpoints (Figure 4 and further described below). Effects were grouped into categories such as carcinogenic, neoplastic, and non-neoplastic pathology, organ weight, etc. As mentioned before, this categorization resulted in terminology for endpoints reported in the studies did not match the terminology in the corresponding guidelines and specifications. This was problematic for two primary reasons: (1) identifying the correct endpoint within a guideline is required to determine whether or not it was negative; and, (2) the endpoint terminology relied on determination of the contribution of a given endpoint to a non-neoplastic or neoplastic process rather than allowing the user to define what effects might be related to cancer phenotypes or other adverse outcomes.

The terminology for both endpoints and effects was standardized to better reflect the terminology used in both the OCSPP guidelines as well as what was reported in the summaries in DERs. The primary change made in ToxRefDB v2 is for the endpoint category “systemic” where the tissue pathology endpoint types are now “pathology microscopic” and “pathology gross,” with no *a priori* suggestion of whether the observation relates to specific cancer or non-cancer related adverse outcomes. Further, duplicative endpoints were standardized, reducing the number of endpoints from approximately 500 to nearly 400. The number of effects remained the same as they were re-binned into the most relevant endpoint. Though the endpoint and effect terminology in ToxRefDBv2 is not comprehensive for all *in vivo* toxicity studies, it captures the observations from the studies and study types currently within ToxRefDB. Each treatment group effect can be further qualified to include life stage, direction of effect (increase, decrease, neutral), target site, and exact terms from the source document used to capture the effect (a field called “effect description free”) (Figure 4).

The endpoint category “neurological” was not updated and has been left out of the release of ToxRefDB v2. The corresponding effects for that endpoint are associated with 18 NEU and 185 DNT studies. However, the study design, dosing, and treatment group information is still available for these studies in the current 2014 release of ToxRefDB v1.3 [25]. The neurological terminology is still under development, with the intention to extend the controlled terminology and extract this information to be available in future updates. The current ToxRefDB v2 controlled terminology is available as Supplemental File 1. The terminology mapping from ToxRefDB v1 to ToxRefDB v2 is available as Supplemental File 2.

2.3.2 Enabling semantic interoperability—Adopting a controlled terminology for ToxRefDB is beneficial for data extraction and data retrieval, but we can also extend the use to enable semantic interoperability across similar resources archiving *in vivo* toxicology data. This will allow interoperability with other sources that also capture *in vivo* toxicology data like Chemical Effects of Biological Systems (CEBS) [44], International Uniform Chemical Information Database (IUCLID) [45], and eTox [46]. Resources such as CDISC actively maintain and update controlled vocabularies for all aspects of nonclinical studies (www.cdisc.org/standards/terminology). Specifically, we were interested in the terminology developed in Standards for Exchange of Nonclinical Data (CDSIC-SEND) and Study Data Tabulation Model (CDISC-SDTM). These vocabularies are maintained by National Cancer Institute Thesaurus (NCIt), which is a subset of the National Library of Medicine’s (NLM) Unified Medical Language System (UMLS) [47]. UMLS is a semantic network linking over 150 terminology resources (CDISC being one of those resources) within the biomedical domain. A UMLS concept is uniquely identified by a concept code. These concept codes were mapped to the controlled terminology defined in ToxRefDB on a manual basis, using the UMLS Terminology Services and NCI Thesaurus browsers. Figure 5 describes the completeness of the mapping for endpoints and effects and the coverage from CDISC-SEND and CDISC-SDTM. The CDISC terminologies were generally applicable to the endpoint level with other terminologies in UMLS needed to map more effects in ToxRefDB, particularly those effects related to developmental and reproductive outcomes. However, as CDISC terminologies are enhanced, the CDISC terms could be automatically mapped utilizing the UMLS codes that have been mapped in this first attempt at including controlled terminology in ToxRefDB. By cross-referencing ToxRefDB terminology with UMLS, a crosswalk to any other resources that adhere to any of the terminology resources maintained within UMLS is enabled. The full endpoint and effect mapping is available as Supplemental File 1.

2.4 Quantitative data extraction and efforts to reduce error rate and data quality

2.4.1 Study extraction process—Initially, ToxRefDB v1 provided only summary effect levels and lacked quantitative dose response information. The quantitative information and its application in ToxRefDB v2 is described in the next section in more detail but served as a strong impetus to re-extract the studies. This task initially proceeded using an Excel file-based extraction; however, the process required manual corrections after uploading study extractions to the ToxRefDB MySQL database, including inconsistent comments, different number of animals for the same treatment group, and added effects outside of the controlled terminology. As a result, an Access database file was generated from the MySQL database for each study (Figure 6), which offers several improvements including standardized options for more consistent reporting in some fields, such as the units on time and dose, dose-treatment group, and effect information; checkbox reporting for observation status on each endpoint and effect; and a log for tracking changes and facilitating QA. Nearly 32% of the studies were extracted using the Excel-based approach, with the remaining studies extracted using the Access database approach. An example Access database file is available as Supplemental File 3; a data dictionary that defines all tables and fields, including those populated using the Access database files, is provided within Supplemental File 4, the ToxRefDB User Guide. Switching to Access database files significantly reduced errors and increased standardization of reporting items such as units, endpoints, and effects.

Guidance for data extraction was stratified first according to study type (e.g., CHR, SUB, DEV, MGR) then by study source (e.g., OPP DER and NTP) because of the differences in both study design and adverse effects required for reporting as stated in guidelines. The process used to extract study information was also an important aspect of QA efforts for ToxRefDB v2. First, a primary reviewer extracted study, dose, treatment group, effect, and endpoint observation information, per standard operating procedures provided to the reviewer. The instructions detailed how to review the toxicological data and extract it from the original data sources consistently across reviewers using the Access database. This was reviewed by a second, senior reviewer, who was asked to review all extracted information as if they were extracting it again and, also, to review the comment log from the primary reviewer. Finally, if either the primary or secondary reviewer noted that it was necessary, an additional senior toxicologist reviewed the comment logs, extracted information, and resolved any conflicts or questions prior to finalization of the extraction. The final, tertiary review occurred for approximately 10% of the studies. All reviewers were trained in the procedures prior to reviewing studies. For release of ToxRefDBv2, the full quantitative data extraction for all CHR and SUB studies were completed, with quantitative data extraction completed for many other study types and sources as well (additional quantitative data will be added in updates to the version 2 release). Table 1 lists the current number of studies with quantitative data extracted by study type and source.

2.4.2 Critical Effect Determination—ToxRefDB v1 and v2 have several effect levels stored, including treatment-related effects that define lowest effect levels (LELs) and no effect levels (NELs); these effects are simply treatment-related and significantly different from control. In contrast, toxicological opinion informs selection of a critical effect designation to define the lowest observable adverse effect and no observable adverse effect levels (LOAELs, NOAELs). A critical effect level is defined as the dose at which a treatment-related effect is deemed to have toxicological significance. Critical effects are typically used to define the study-level POD for regulatory toxicology applications. Not all studies within ToxRefDB v1 had been assessed for critical effects, or the critical effects had not been extracted. For extractions of OPP DER files, the critical effect was simply captured from the source document as previously identified by toxicologists who had reviewed the original study file. If critical effect information was lacking for a given source document, i.e., those from sources other than OPP DER files including preclinical studies from the pharmaceutical industry, NTP study files, and the open literature, senior toxicologists trained to extract information for ToxRefDB reviewed the study and determined the critical effects and critical effect levels. For each study from one of these non-OPP DER sources, the reviewers determined the critical effect and lowest observed adverse effect level(s) (LOAEL) using a weight-of-evidence (WoE) approach, [48], similar to the approach used to evaluate registrant-submitted studies for generation of DERs. Using this approach, the identification of potential critical effects from a given study was determined based on statistical significance, considerations of biological relevance, and consistency across multiple endpoints (in the presence or absence of statistical significance) to select the appropriate LOAEL value(s) and the overall study LOAEL. The WoE evaluation included review of all pertinent information so that the full impact of biological plausibility and coherence was adequately considered. This approach involves weighing individual lines of evidence and combining the entire body of evidence to make an informed judgment. Judgment about the WoE involved considerations of the quality and adequacy of data, and consistency of responses induced by the agent in question. The WoE judgment required combined input of relevant disciplines. Generally, no single factor determined the overall weight; all potential factors were judged in combination. The results of these reviews were recorded along with appropriate rationales and can be found in ToxRefDB v2.

2.4.3 Quality assurance and quality control efforts to reduce error rate—Error rate is an inherent problem for legacy databases as much of the source information was entered manually and human errors resulting from transcription are impossible to completely avoid [31]. However, as part of the ToxRefDB v2 effort, more robust QA processes were implemented to promote greater fidelity of the information extracted and numerous quality control (QC) checks to verify data integrity. First, studies were extracted utilizing a defined QA process, with multiple levels of review and Access form-based entry (described previously) to prevent extraction errors. Upon upload into ToxRefDB v2, these extractions were required to pass specific QC checks because, although the Access database files enforce the MySQL database constraints as well as use of the controlled terminology to minimize data entry error, logical errors can persist. We checked a series of potential logical errors after the extracted data was uploaded through the import script. These errors were identified by defining a series of tests up front that must resolve to a particular answer. Below are some of the logical errors that were flagged using a QC check following import of the Access database files to the MySQL database:

- Dose level numbering did not correspond to the total number of doses;
- Duplication of concentration/dose values, including two control doses;
- No concentration and no dose adjusted value for a reported effect (possible extraction error or possibly that the effect was qualitatively reported);
- The critical effect level is at a dose below where treatment-related effects were observed; and/or,
- The control was incorrectly identified as a critical effect level.

Any of these issues that could not be resolved systematically were flagged to undergo a second round of extraction and review to correct. Though QC is an ongoing and evolving process, these QC checks are serving as an improvement to the overall database and database development process.

An additional ongoing problem for reporting quantitative data from clinical or related laboratory findings is unit standardization [49, 50]. No guidance is provided on how to report findings in the OCSPP guidelines nor from any other sources, so units were extracted exactly as they were presented in the reports. The units were standardized by eliminating duplicate entries for the same units that were originally entered differently or with typographical errors. Units were only standardized and no conversions were introduced in the current database. Further work must be undertaken to further standardize units and define conversions that can be systematically automated. The current progress for this work is available in Supplemental File 5 where both the original and corrected units are provided.

In order to understand how increased QA and QC may have affected quantitative information in ToxRefDB, a comparison of study level LEL and LOAEL values for 3,446 studies between ToxRefDB v1.3 and v2 was conducted. This evaluation showed ~95% concordance for the LELs and ~90% for LOAELs between ToxRefDB v1.3 and ToxRefDB v2. The summary effect level values in v1.3 and v2 were largely concordant. The magnitude of the differences in these values made by error correction ranged from 0.1 log10-mg/kg/day to 2.4 log10-mg/kg/day, with an average difference of 0.52 and 0.57 log10-mg/kg/day for LEL and LOAELs, respectively.

2.4.4 Distinctions between negative effects and not tested effects—Many study sources only report information on the adverse effects, and data extracted in ToxRefDB v1 reflected this practice (i.e. only contained data for positive or treatment-related effects). These values were reported as lowest effect levels (LELs) or lowest observable adverse effect levels (LOAELs), with no effect and no observable adverse effect levels (NELs, NOAELs) inferred as the next lowest dose, respectively. A positives-only database presented a major challenge for predictive modeling applications that require balanced training sets of positive and negative findings because the user was left to infer negatives from the database without the guidance of what was tested and reported for the study based on its adherence (or non-adherence) to a guideline. Finding a solution to systematic and accurate inference of negatives involved leveraging the new controlled effect terminology to match the OCSPP guidelines (described above) and annotating endpoints as required, triggered, or recommended. Required endpoints are always tested according to the guidelines, whereas triggered endpoints are required under specific circumstances, e.g. if a chemical is known to perturb a specific system based on information from previous studies. A recommended endpoint is not always tested but are mentioned as important in the guidelines. All other endpoints not explicitly mentioned in the guidelines were assumed to be not required. The collections of endpoint annotations for guidelines are referred to as guideline profiles. All guideline profiles developed for ToxRefDB v2 are available as Supplemental File 6.

These guideline profiles enable assumptions about whether an endpoint was tested for a given study based on which guideline the study followed. A majority of the studies (58%) described in ToxRefDB are based on OPP DERs, which summarize registrant-submitted data for OCSPP series 870 Health Effects Testing Guidelines as seen in Table 2. Additionally, though not strictly referred to as guidelines, study specifications for SAC, SUB, and CHR studies from NTP [51] were also reviewed and developed into guideline profiles to allow for their inclusion in determination of negatives. Developmental and reproductive studies from the NTP were not included in guideline profile development at this time due to the assumption that these studies may have been highly customized based on the experimental need, and as such inference of negatives may not lead to accurate conclusions because the endpoints and effects tested between studies may be variable (personal communication, John Bucher and Paul Foster). Because the studies included in ToxRefDB span decades, we also included guideline profiles for updated guidelines. For example, since testing requirements were added to the MGR guideline (OCSPP 870.3800) in 1998 [33], the MGR study type has two associated guideline profiles: one for studies conducted before 1998 and another for studies conducted in 1998 and later. All of the guideline profiles were reviewed by an independent senior toxicologist familiar with the guideline and guidance documents.

Observations were recorded and confirmed in the data extraction process for each study to reflect concordance with guideline profiles, deviations, endpoints that were measured following a trigger, etc. An observation is defined as the testing and reporting status of a given endpoint in the study. Data extractors made decisions about testing and reporting status as described in Table 3, where for example endpoints that were reported as tested can be differentiated from endpoints that are assumed to be tested based on the guideline profile. The important result of the development of these guideline profiles is that missing or not tested data can now be distinguished from negative (tested with no effect seen) for a large fraction of the studies described in ToxRefDB v2. The inference workflow to determine negative effects based on observations and guideline profiles is described in Fig. 7.

2.4.5 Study Reliability (ToxRTool)—A majority of the studies referenced within ToxRefDB were extracted via summaries from OPP DERs, and these studies typically follow OCSPP 870 series Health Effects Testing Guidelines; however, as ToxRefDB was expanded, other studies were summarized from various sources, including: NTP, pharmaceutical companies, published literature (OpenLit), and other. NTP and pre-clinical pharmaceutical studies were considered guideline-like, as a study guideline or specification that these studies resembled could be identified, but OpenLit studies were not assumed to conform to any guideline. Therefore, all OpenLit studies were assessed for reliability and guideline adherence. The Toxicological Data Reliability Assessment Tool (ToxRTool) was adapted for reliability assessment [52]. ToxRTool is an Excel application that includes questions across 5 criteria with numerical responses that are summed to lead to a Klimisch score: a score ranging from 1-4 that captures an overall assessment of reliability [53]. The ToxRTool was adapted specifically in the following ways for this project:

- Added Guideline Adherence Score (an initial question for the reviewer regarding the study’s adherence to or consistency with OCSPP guidelines with a five-point rating scale) further described in Table 4.
- Added “Context of Tool and Rationale/Intent for Study” field [an open-text field to insert the purpose of the study quality review to address the concern raised by Segal et al. (2015) [54] that the intended purpose of the ToxRTool-facilitated review could influence evaluations].
- Added additional scoring notes (to help the reviewers assign scores consistently).
- Added option for “0.5” rating for selected criteria (for some questions considered more subjective than others, if the reviewer concluded the question was partially fulfilled).

A total of 522 OpenLit studies were assessed with the ToxRTool with scores ranging from 8 to 23. The majority of the studies reviewed for ToxRefDB v2 corresponded to Klimisch quality scores of 1 (ToxRTool score of ≥ 18) or 2 (ToxRTool score of 13-18). The ToxRTool scores could be used as a quality flag both to qualify and prioritize studies for the extraction process, or by users who are performing reviews of information on a single chemical basis.

2.5 ToxRefDB v2 utility for research applications

The following sections indicate a few examples of how improvements to ToxRefDB v2 may facilitate research applications. ToxRefDB v2 is a research product that can be useful for asking predictive and computational toxicology questions and is not meant to be a database for understanding the regulatory status or decisions for any given substance. The ToxRefDB User Guide (Supplemental File 4) provides detailed information including a data dictionary to describe details of the tables and fields in the database, and example code to retrieve information from the database for common research questions.

2.5.1 Systematic calculation of point of departures (PODs) and related effect levels—Related to the new ToxRefDB v2 controlled effect terminology is the application of this terminology for calculation of PODs and related effect levels for various modeling and retrospective analyses. For purposes of predictive toxicology, PODs can be computed per chemical (i.e., lowest dose that produced effects or adverse effects across all study types included in the database) or per study (i.e., lowest dose that produced effects or adverse effects in a given study of interest). PODs computed by chemical could be broken down into a POD for some combination of effects in a POD “category,” e.g., the lowest dose that produced effects or adverse effects on developmental or reproductive effects as a group. Acknowledging that the specific application may define the appropriate aggregation of the effect data in ToxRefDB for calculation of PODs, ToxRefDB v2 (Figure 3) enables definition of the list of effects to be grouped together, followed by storage of the PODs calculated based on that list. A collection of effect groupings is referred to as an effect profile. An initial set of effect profiles were created to define custom grouping of effects from the study, treatment groups, and effects. For example, all developmental effects, across studies, could be combined to give a POD, or minimum LOAEL or LEL value, for developmental effects. The NEL and NOAEL are designated as the next lowest doses from the LEL and LOAEL, respectively. A complication in providing PODs is that not every effect is necessarily of toxicological significance and may not correspond to the critical effect level as reviewed by toxicologists. In the case that no effects of toxicological significance were observed for a given category of effects, or by study, the LOAEL is greater than the highest dose tested and the NOAEL is greater than or equal to the highest dose tested for that effect (i.e., a "free-standing NOAEL"). For all POD types, including NEL, NOAEL, LEL, and LOAEL, a qualifier (<, >, or =) is provided to assist with quantitative interpretation of these values. For example, if a NOAEL is equal to the highest dose tested, and the LOAEL is greater than the highest dose tested, then the study did not precisely indicate the threshold dose for a given effect profile.

The effect profiles are an important feature addition and address problems previously highlighted [28]; essentially, the endpoints and effects in ToxRefDB can be grouped a number of ways, which may lead to differing interpretations. However, there is no single way to create POD values via grouping of effects, as differing interpretations may be equally valid for divergent applications of the data. The two effect profiles currently available in ToxRefDB v2 are summarized in Table 5 for clarity, with the expectation that as use of the database grows, additional effect profiles can be added. It should be noted that these effect profiles, and the POD values generated in using them, are for research purposes and do not necessarily reflect POD values that may be used in chemical safety evaluations.

First, effect level data were grouped by study type, endpoint category, and life stage (available as Supplemental File 7). This first effect profile produced POD values for each study type, life stage, and endpoint category combination. This first effect profile was used to calculate effect levels for the CompTox Chemicals Dashboard [55].

A second effect profile was also employed (also available as Supplemental File 7), where PODs were calculated for each endpoint category-endpoint type pairing, except in the case of the systemic endpoint category, where PODs were reported for each endpoint target (i.e., organs). This second effect profile produced POD values for cholinesterase, developmental, and reproductive endpoint categories; hematology, in-life observation, and urinalysis endpoint types; and organ-specific endpoint targets (e.g., liver). Either of these effect groupings and associated effect levels may be useful for research purposes as a meaningful way of considering many pieces of information for a chemical at one time.

2.5.2 Quantitative data and benchmark dose (BMD) modeling for ToxRefDB—

In this major update, there are now quantitative data available in ToxRefDBv2 (as described in Section 2.4.4), with dose, effect, treatment group size, and where available, error estimates (e.g. standard deviation, standard error) provided for each treatment group. These data are necessary for dose-response modeling. Quantitative dose-response modeling yields

PODs for research applications that are less dependent on the doses selected for a given study, and ensures that dose values selected correspond to similar levels of effects across studies [56]. Though there are many possible approaches to curve-fitting [57, 58], the US EPA Benchmark Dose Modeling Software (BMDS) [59–61] has become the canonical tool for use in toxicology regulatory and research applications [58, 62–64]. Using BMDS to fit the quantitative response data in ToxRefDB provides modeled values, e.g., benchmark dose (BMD) values, using the default recommendations from the BMDS guidance [30]. In ToxRefDB v2, we report the results from the largest use of batch processing with BMDS v2.7 employed to date, using a Python wrapper [65]. The technical specifications of the Python-driven BMDS are reported in an accompanying application note [65], and the associated code for processing ToxRefDB v2 and storing the BMD values is reported in Supplemental Files 8–10. The objectives in reporting this demonstration within the database are (1) to enable and promote the use of BMD values in predictive toxicology instead of the tested doses; and (2) to demonstrate the feasibility of large-scale BMDS analysis of legacy toxicology information. The BMD values reported are not intended to reflect any regulatory decision-making on a single chemical basis.

Over 92,000 quantitative dose-response datasets, i.e. data from chemical-effect pairs, from complete study extractions with at least 3 non-control dose levels in ToxRefDB v2 were filtered to yield datasets amenable to modeling using BMDS. For each dose group of a study, BMDS analysis requires the dose, N, and dichotomous incidence or the continuous effect level mean and variance. In large part due to inconsistent or incomplete reporting of variance for continuous responses, only about one-third of the total datasets were amenable for BMDS modeling (nearly 28,000). For each modeled dose-response, the data were grouped according to the response type, i.e., continuous response, continuous response for organ and body weights, dichotomous response, or dichotomous cancer response. The response type guided selection of the models and benchmark response (BMR) used in the automated analysis, as shown in Table 6, as recommended by the BMDS guidance [30]. The effect terminology corresponding to cancer is available in Supplemental File 11.

The current BMD table in ToxRefDB v2 holds results for nearly 28,000 datasets. This includes BMD models with 1 standard deviation for continuous data, 10% relative deviation for organ/body weight, and 5 and 10% BMR for dichotomous data. Almost 90% of the datasets were successfully modeled and have at least one recommended model and associated BMD value (Figure 8). Currently, there are 627 unique chemicals (as indicated by CAS registry number) with at least one modeled BMD value. The lower 95% confidence limit on the BMD value estimate, known as the BMDL, is also stored in the database. However, some recommended models are associated with cautions for using the BMD and BMDL values that were auto-generated by BMDS (v2.7). For example, there may be warnings to indicate a large distance between the BMD and BMDL, or that a computed BMDL is likely imprecise because the model has not converged. All warnings related to model recommendations are stored in the “logic cautions” column of the “bmd” table in the database and should be considered by the end user of the data. The fraction of chemical-effect data for which a BMD value could be generated (with a recommended model achieved) for each data type and BMD model type are illustrated in Figure 9.

One hypothesis in modeling these dose-response data is that BMDL values will tend to be lower than the discrete NOAEL or NEL values for a given study-level effect. Indeed, most of the recommended BMDL values are less than the stored NOAEL and NEL values for that effect. For datasets (continuous and dichotomous) successfully modeled using a BMR of 5% extra risk (dichotomous) or 5% increase relative to control mean estimate (continuous), 99.6% of the BMDL values were less than the corresponding NOAEL values, and 72% of the BMDL values were less than the corresponding NEL values. For the datasets successfully modeled using a BMR of 10% extra risk (dichotomous) or 10% increase relative to control mean estimate (continuous), 99.6% of the BMDL values were lower than the NOAEL values, and 52% of the BMDLs were lower than the corresponding NEL values. For the datasets successfully modeled using a BMR of 1 standard deviation (continuous data only), 100% of the BMDL were less than the corresponding NOAEL values, and 46% of the BMDL values were lower than the corresponding NEL values. This is mostly consistent with previous works showing that modeled BMDLs are a more conservative estimation of PODs than the statistically derived NOAELs [66, 67]. Similarly, we also compared BMD values to LEL and LOAEL values. For the datasets successfully modeled using a BMR of 5% extra risk (dichotomous) or 5% increase relative to control mean estimate (continuous), BMD values were less than their corresponding LEL or LOAEL values approximately 91% of the time. For the datasets successfully modeled using a BMR of 10% extra risk (dichotomous) or 10% increase relative to control mean estimate (continuous), the BMD values were less than the corresponding LEL or LOAEL values approximately 81% of the time. For the datasets successfully modeled using a BMR of 1 standard deviation (continuous data only), the BMD values were less than the corresponding LEL or LOAEL values 80% and 81% of the time, respectively. These calculations are provided in Supplemental File 12. In general, these findings suggest that the BMDL and BMD tend to be lower dose estimates of NEL/NOAEL and LEL/LOAEL, respectively.

All study type and BMDS-amenable data were used for this exercise to model the largest dataset possible. The BMDS results reported within ToxRefDB v2 are intended to provide input values for research and modeling applications. Some caution needs to be taken when evaluating the BMD models, particularly for studies where multiple generations are evaluated. Ideally, for MGR and DEV studies, a nested model should be used to calculate BMDs for the litters and, if needed, a correction for the degree of variability or sample size adjustment. However, due to the availability of information from source files, and the data structure in ToxRefDB v2, information from individuals in each litter were not available. Therefore, the summary data and statistics for litters were used for BMD modeling.

2.5.3 Support for Data Integration—It is increasingly apparent that many toxicology research questions will require the integration of public data resources, both with those containing the same types of information, as well as with other databases to connect different kinds of information. For example, efforts linking *in vitro* effects in ToxCast to *in vivo* outcomes using predictive models may help to identify rapid, more efficient chemical screening alternatives [68]. To connect the ToxRefDB endpoint and effect terminology with other resources, the ToxRefDB terminology was standardized and cross-referenced to the United Medical Language System (UMLS). UMLS cross-references enable mapping of *in* *vivo* pathological effects from ToxRefDB to PubMed (via Medical Subject Headings or MeSH terms), which may be relevant for toxicological research and systematic review. This enables linkage to any resource that is also connected to PubMed or indexed with MeSH. For example, Entity MeSH Co-occurrence Network (EMCON) [69], a resource to retrieve ranked lists of genes for a given topic, can be used to identify genes related to adverse effects observed in ToxRefDB. Subsequently, ToxCast can be integrated since the intended targets are mapped to Entrez gene IDs. The result of updating the terminology in ToxRefDB v2 and linking to the UMLS concepts is that ToxRefDB may be used to better anchor or compare to NAM data like ToxCast data, or to other *in vivo* databases of toxicological information, like those available from eChemPortal [70], e-TOX [46], or others. Integration of these data resources is a major hurdle toward to evaluating the reproducibility and biological meaning of both traditional, legacy toxicity information and the data from NAMs.

3 Conclusions

ToxRefDB has served as a seminal resource for in vivo toxicity studies with broad applications in predictive modeling, retrospective analysis, and evaluation of NAM results. Although robust in scope for capturing effect information, early versions of ToxRefDB only contained data for positive findings and thus were limited by a lack of distinction between tested and not tested effects. Additionally, the terminology concerning endpoints did not adhere to a standardized classification system. Moreover, specific effect information and quantitative, dose-response information were needed to support a broader range of predictive toxicology applications. To address these issues, ToxRefDB has undergone extensive updates that include extraction of additional information (quantitative data as well as observations about tested endpoints), data standardization, and QA and QC measures to maintain data integrity. With these updates, the utility of ToxRefDB can be extended to myriad applications, and our process can serve as a reference for other resources aggregating similar information. The features added in this release of ToxRefDB v2 support ongoing efforts to use these data to train predictive models and also to evaluate the reproducibility and variability in existing animal-based approaches for safety testing used for model training and performance evaluation. The MySQL database and all associated summary flat files are available at ftp://newftp.epa.gov/comptox/High_Throughput_Screening_Data/Animal_Tox_Data/current/. Further documentation, code, and examples are available at https://github.com/USEPA/CompTox-ToxRefDB.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

The authors would like to acknowledge Josephine Brown, Angelica Adrian, Nathaniel Rush, and Jason Brown for contributions over the course of redevelopment of ToxRefDB; Andy Shapiro formerly at the National Toxicology Program and currently at Infinia ML for development of and assistance with python-bmds; Jennifer Fostel at the National Toxicology Program for ongoing discussions on integrating legacy data resources; Bryan Luukinen at Duke University Superfund Research Center for management and training for data extraction; and Joanne Trgovcich at ICF for coordinating the ToxRTool work. The authors would also like to thank Scott Auerbach at the National Toxicology Program and Jeff Gift, Reeder Sams, Woody Setzer, and Imran Shah (US EPA) for insightful comments on previous versions of this manuscript.

4 Funding

L.L.P. was supported by appointment to the Research Participation Program of the U.S. Environmental Protection Agency, Office of Research and Development, administered by the Oak Ridge Institute for Science and Education through an interagency agreement between the U.S. Department of Energy and the U.S. EPA.

6. References

[1]. Dix DJ, Houck KA, Martin MT, Richard AM, Setzer RW, Kavlock RJ, The ToxCast program for prioritizing toxicity testing of environmental chemicals, Toxicol Sci 95(1) (2007) 5–12. [PubMed: 16963515]

[2]. NRC, Toxicity Testing in the 21st Century: A Vision and a Strategy, National Academics Press, Washington, DC, 2007.

[3]. Collins FS, Gray GM, Bucher JR, Toxicology. Transforming environmental health protection, Science 319(5865) (2008) 906–7. [PubMed: 18276874]

[4]. Martin MT, Judson RS, Reif DM, Kavlock RJ, Dix DJ, Profiling chemicals based on chronic toxicity results from the U.S. EPA ToxRef Database, Environ Health Perspect 117(3) (2009) 392–9. [PubMed: 19337514]

[5]. Martin MT, Mendez E, Corum DG, Judson RS, Kavlock RJ, Rotroff DM, Dix DJ, Profiling the reproductive toxicity of chemicals from multigeneration studies in the toxicity reference database, Toxicol Sci 110(1) (2009) 181–90. [PubMed: 19363143]

[6]. Kleinstreuer NC, Dix DJ, Houck KA, Kavlock RJ, Knudsen TB, Martin MT, Paul KB, Reif DM, Crofton KM, Hamilton K, Hunter R, Shah I, Judson RS, In vitro perturbations of targets in cancer hallmark processes predict rodent chemical carcinogenesis, Toxicol Sci 131(1) (2013) 40–55. [PubMed: 23024176]

[7]. Sipes NS, Martin MT, Reif DM, Kleinstreuer NC, Judson RS, Singh AV, Chandler KJ, Dix DJ, Kavlock RJ, Knudsen TB, Predictive models of prenatal developmental toxicity from ToxCast high-throughput screening data, Toxicol Sci 124(1) (2011) 109–27. [PubMed: 21873373]

[8]. Theunissen PT, Beken S, Cappon GD, Chen C, Hoberman AM, van der Laan JW, Stewart J, Piersma AH, Toward a comparative retrospective analysis of rat and rabbit developmental toxicity studies for pharmaceutical compounds, Reprod Toxicol 47 (2014) 27–32. [PubMed: 25517003]

[9]. Martin MT, Knudsen TB, Reif DM, Houck KA, Judson RS, Kavlock RJ, Dix DJ, Predictive model of rat reproductive toxicity from ToxCast high throughput screening, Biol Reprod 85(2) (2011) 327–39. [PubMed: 21565999]

[10]. Knudsen TB, Martin MT, Kavlock RJ, Judson RS, Dix DJ, Singh AV, Profiling the activity of environmental chemicals in prenatal developmental toxicity studies using the U.S. EPA's ToxRefDB, Reprod Toxicol 28(2) (2009) 209–19. [PubMed: 19446433]

[11]. Liu J, Mansouri K, Judson RS, Martin MT, Hong H, Chen M, Xu X, Thomas RS, Shah I, Predicting hepatotoxicity using ToxCast in vitro bioactivity and chemical structure, Chem Res Toxicol 28(4) (2015) 738–51. [PubMed: 25697799]

[12]. Thomas RS, Black MB, Li L, Healy E, Chu TM, Bao W, Andersen ME, Wolfinger RD, A comprehensive statistical analysis of predicting in vivo hazard using high-throughput in vitro screening, Toxicol Sci 128(2) (2012) 398–417. [PubMed: 22543276]

[13]. Novotarskyi S, Abdelaziz A, Sushko Y, Korner R, Vogt J, Tetko IV, ToxCast EPA in Vitro to in Vivo Challenge: Insight into the Rank-I Model, Chem Res Toxicol 29(5) (2016) 768–75. [PubMed: 27120770]

[14]. Truong L, Ouedraogo G, Pham L, Clouzeau J, Loisel-Joubert S, Blanchet D, Nocairi H, Setzer W, Judson R, Grulke C, Mansouri K, Martin M, Predicting in vivo effect levels for repeat-dose systemic toxicity using chemical, biological, kinetic and study covariates, Arch Toxicol 92(2) (2018) 587–600. [PubMed: 29075892]

[15]. Judson RS, Martin MT, Patlewicz G, Wood CE, Retrospective mining of toxicology data to discover multispecies and chemical class effects: Anemia as a case study, Regul Toxicol Pharmacol 86 (2017) 74–92. [PubMed: 28242142]

[16]. Hill T 3rd, Nelms MD, Edwards SW, Martin M, Judson R, Corton JC, Wood CE, Editor's Highlight: Negative Predictors of Carcinogenicity for Environmental Chemicals, Toxicol Sci 155(1) (2017) 157–169. [PubMed: 27679563]

[17]. Casati S, Aschberger K, Barroso J, Casey W, Delgado I, Kim TS, Kleinstreuer N, Kojima H, Lee JK, Lowit A, Park HK, Regimbald-Krnel MJ, Strickland J, Whelan M, Yang Y, Zuang V, Standardisation of defined approaches for skin sensitisation testing to support regulatory use and international adoption: position of the International Cooperation on Alternative Test Methods, Arch Toxicol 92(2) (2018) 611–617. [PubMed: 29127450]

[18]. Covell DG, Integrating constitutive gene expression and chemoactivity: mining the NCI60 anticancer screen, PLoS One 7(10) (2012) e44631. [PubMed: 23056181]

[19]. Zhao F, Li R, Xiao S, Diao H, El Zowalaty AE, Ye X, Multigenerational exposure to dietary zearalenone (ZEA), an estrogenic mycotoxin, affects puberty and reproduction in female mice, Reprod Toxicol 47 (2014) 81–8. [PubMed: 24972337]

[20]. Sutton P, Woodruff TJ, Perron J, Stotland N, Conry JA, Miller MD, Giudice LC, Toxic environmental chemicals: the role of reproductive health professionals in preventing harmful exposures, Am J Obstet Gynecol 207(3) (2012) 164–73. [PubMed: 22405527]

[21]. Fourches D, Muratov E, Tropsha A, Trust, but verify: on the importance of chemical structure curation in cheminformatics and QSAR modeling research, J Chem Inf Model 50(7) (2010) 1189–204. [PubMed: 20572635]

[22]. Kavlock R, Chandler K, Houck K, Hunter S, Judson R, Kleinstreuer N, Knudsen T, Martin M, Padilla S, Reif D, Richard A, Rotroff D, Sipes N, Dix D, Update on EPA's ToxCast program: providing high throughput decision support tools for chemical risk management, Chem Res Toxicol 25(7) (2012) 1287–302. [PubMed: 22519603]

[23]. Tice RR, Austin CP, Kavlock RJ, Bucher JR, Improving the human hazard characterization of chemicals: a Tox21 update, Environ Health Perspect 121(7) (2013) 756–65. [PubMed: 23603828]

[24]. Richard AM, Judson RS, Houck KA, Grulke CM, Volarath P, Thillainadarajah I, Yang C, Rathman J, Martin MT, Wambaugh JF, Knudsen TB, Kancherla J, Mansouri K, Patlewicz G, Williams AJ, Little SB, Crofton KM, Thomas RS, ToxCast Chemical Landscape: Paving the Road to 21st Century Toxicology, Chem Res Toxicol 29(8) (2016) 1225–51. [PubMed: 27367298]

[25]. NCCT, Animal Toxicity Studies: Effects and Endpoints (Toxicity Reference Database - ToxRefDB files), 2018.

[26]. NCCT, ReadMe for Animal Toxicity Study ToxRefDB files, 2018.

[27]. Janus E, Concerns of CropLife America Regarding the Application and Use of the U.S. EPA's Toxicity Reference Database, Environmental Health Perspectives 117(10) (2009) A432–A432.

[28]. Plunkett LM, Kaplan AM, Becker RA, Challenges in using the ToxRefDB as a resource for toxicity prediction modeling, Regul Toxicol Pharmacol 72(3) (2015) 610–4. [PubMed: 26003516]

[29]. Liu J, Patlewicz G, Williams AJ, Thomas RS, Shah I, Predicting Organ Toxicity Using in Vitro Bioactivity Data and Chemical Structure, Chem Res Toxicol 30(11) (2017) 2046–2059. [PubMed: 28768096]

[30]. E.P.A. U.S, Benchmark dose technical guidance, U.S. Environmental Protection Agency, Risk Assessment Forum, Washington, DC, 2012.

[31]. Wahi MM, Parks DV, Skeate RC, Goldin SB, Reducing errors from the electronic transcription of data collected on paper forms: a research data case study, J Am Med Inform Assoc 15(3) (2008) 386–9. [PubMed: 18308994]

[32]. Majid A, Bae ON, Redgrave J, Teare D, Ali A, Zemke D, The Potential of Adaptive Design in Animal Studies, Int J Mol Sci 16(10) (2015) 24048–58. [PubMed: 26473839]

[33]. USEPA, Health Effects Test Guidelines: OPPTS 870.3800 Reproduction and Fertility Effects 1998.

[34]. Courtot M, Juty N, Knupfer C, Waltemath D, Zhukova A, Drager A, Dumontier M, Finney A, Golebiewski M, Hastings J, Hoops S, Keating S, Kell DB, Kerrien S, Lawson J, Lister A, Lu J, Machne R, Mendes P, Pocock M, Rodriguez N, Villeger A, Wilkinson DJ, Wimalaratne S, Laibe

C, Hucka M, Le Novere N, Controlled vocabularies and semantics in systems biology, Mol Syst Biol 7 (2011) 543. [PubMed: 22027554]

[35]. Ward JM, Schofield PN, Sundberg JP, Reproducibility of histopathological findings in experimental pathology of the mouse: a sorry tail, Lab Anim (NY) 46(4) (2017) 146–151. [PubMed: 28328876]

[36]. Wolf JC, Maack G, Evaluating the credibility of histopathology data in environmental endocrine toxicity studies, Environ Toxicol Chem 36(3) (2017) 601–611. [PubMed: 27883231]

[37]. STP, International Harmonization of Nomenclature and Diagnostic Criteria (INHAND) Published Guides, 2018 https://www.toxpath.org/inhand.asp#pubg (Accessed August 2018.

[38]. Moreno-Conde A, Moner D, Cruz WD, Santos MR, Maldonado JA, Robles M, Kalra D, Clinical information modeling processes for semantic interoperability of electronic health records: systematic review and inductive analysis, J Am Med Inform Assoc 22(4) (2015) 925–34. [PubMed: 25796595]

[39]. Evans RS, Electronic Health Records: Then, Now, and in the Future, Yearb Med Inform Suppl 1 (2016) S48–61.

[40]. Kuchinke W, Aerts J, Semler SC, Ohmann C, CDISC standard-based electronic archiving of clinical trials, Methods Inf Med 48(5) (2009) 408–13. [PubMed: 19621114]

[41]. Kaufman L, Gore K, Zandee JC, Data Standardization, Pharmaceutical Drug Development, and the 3Rs, ILAR J 57(2) (2016) 109–119. [PubMed: 28053065]

[42]. Cesta MF, Malarkey DE, Herbert RA, Brix A, Hamlin MH 2nd, Singletary E, Sills RC, Bucher JR, Birnbaum LS, The National Toxicology Program Web-based nonneoplastic lesion atlas: a global toxicology and pathology resource, Toxicol Pathol 42(2) (2014) 458–60. [PubMed: 24488020]

[43]. USEPA, Health Effects Test Guidelines: OPPTS 870.4100 Chronic Toxicity, 1998.

[44]. Lea IA, Gong H, Paleja A, Rashid A, Fostel J, CEBS: a comprehensive annotated database of toxicological data, Nucleic Acids Res 45(D1) (2017) D964–D971. [PubMed: 27899660]

[45]. Heidorn CJ, Rasmussen K, Hansen BG, Norager O, Allanou R, Seynaeve R, Scheer S, Kappes D, Bernasconi R, IUCLID: an information management tool for existing chemicals and biocides, J Chem Inf Comput Sci 43(3) (2003) 779–86. [PubMed: 12767136]

[46]. Briggs K, Barber C, Cases M, Marc P, Steger-Hartmann T, Value of shared preclinical safety studies - The eTOX database, Toxicol Rep 2 (2015) 210–221. [PubMed: 28962354]

[47]. Bodenreider O, The Unified Medical Language System (UMLS): integrating biomedical terminology, Nucleic Acids Res 32(Database issue) (2004) D267–70. [PubMed: 14681409]

[48]. Hardy A, Benford D, Halldorsson T, Jeger MJ, Knutsen HK, More S, Naegeli H, Noteborn H, Ockleford C, Ricci A, Rychen G, Schlatter JR, Silano V, Solecki R, Turck D, Benfenati E, Chaudhry QM, Craig P, Frampton G, Greiner M, Hart A, Hogstrand C, Lambre C, Luttik R, Makowski D, Siani A, Wahlstroem H, Aguilera J, Dorne J-L, Fernandez Dumont A, Hempen M, Valtueña Martínez S, Martino L, Smeraldi C, Terron A, Georgiadis N, Younes M, Guidance on the use of the weight of evidence approach in scientific assessments, EFSA Journal 15(8) e04971.

[49]. Simpson D, Units for reporting the results of toxicological measurements, Ann Clin Biochem 17(6) (1980) 328–31. [PubMed: 7212606]

[50]. Zegers I, Schimmel H, To Harmonize and Standardize: Making Measurement Results Comparable, Clinical Chemistry 60(7) (2014) 911. [PubMed: 24799526]

[51]. NTP, Specifications for the conduct of studies to evaluate the toxic and carcinogenic potential of chemical, biological and physical agents in laboratory animals for the national toxicology program (NTP), RTP, NC, 2006.

[52]. Schneider K, Schwarz M, Burkholder I, Kopp-Schneider A, Edler L, Kinsner-Ovaskainen A, Hartung T, Hoffmann S, “ToxRTool”, a new tool to assess the reliability of toxicological data, Toxicol Lett 189(2) (2009) 138–44. [PubMed: 19477248]

[53]. Klimisch HJ, Andreae M, Tillmann U, A systematic approach for evaluating the quality of experimental toxicological and ecotoxicological data, Regul Toxicol Pharmacol 25(1) (1997) 1–5. [PubMed: 9056496]

[54]. Segal D, Makris SL, Kraft AD, Bale AS, Fox J, Gilbert M, Bergfelt DR, Raffaele KC, Blain RB, Fedak KM, Selgrade MK, Crofton KM, Evaluation of the ToxRTool's ability to rate the reliability of toxicological data for human health hazard assessments, Regul Toxicol Pharmacol 72(1) (2015) 94–101. [PubMed: 25777839]

[55]. Williams AJ, Grulke CM, Edwards J, McEachran AD, Mansouri K, Baker NC, Patlewicz G, Shah I, Wambaugh JF, Judson RS, Richard AM, The CompTox Chemistry Dashboard: a community data resource for environmental chemistry, J Cheminform 9(1) (2017) 61. [PubMed: 29185060]

[56]. Haber LT, Dourson ML, Allen BC, Hertzberg RC, Parker A, Vincent MJ, Maier A, Boobis AR, Benchmark dose (BMD) modeling: current practice, issues, and challenges, Crit Rev Toxicol 48(5) (2018) 387–415. [PubMed: 29516780]

[57]. Filer DL, Kothiya P, Setzer RW, Judson RS, Martin MT, tcpl: the ToxCast pipeline for high-throughput screening data, Bioinformatics 33(4) (2017) 618–620. [PubMed: 27797781]

[58]. Hardy A, Benford D, Halldorsson T, Jeger Michael J, Knutsen Katrine H, More S, Mortensen A, Naegeli H, Noteborn H, Ockleford C, Ricci A, Rychen G, Silano V, Solecki R, Turck D, Aerts M, Bodin L, Davis A, Edler L, Gundert-Remy U, Sand S, Slob W, Bottex B, Abrahantes Jose C, Marques Daniele C, Kass G, Schlatter Josef R, Update: use of the benchmark dose approach in risk assessment, EFSA Journal 15(1) (2017) e04658.

[59]. Davis JA, Gift JS, Zhao QJ, Introduction to benchmark dose methods and U.S. EPA's benchmark dose software (BMDS) version 2.1.1, Toxicol Appl Pharmacol 254(2) (2011) 181–91. [PubMed: 21034758]

[60]. E.P.A. U.S, Benchmark Dose Software (BMDS), National Center for Environmental Assessment, Research Triangle Park, NC, 2011.

[61]. E.P.A. U.S, Benchmark Dose Software (BMDS). Version 2.2, 2012.

[62]. Gephart LA, Salminen WF, Nicolich MJ, Pelekis M, Evaluation of subchronic toxicity data using the benchmark dose approach, Regul Toxicol Pharmacol 33(1) (2001) 37–59. [PubMed: 11259178]

[63]. Fournier K, Tebby C, Zeman F, Glorennec P, Zmirou-Navier D, Bonvallot N, Multiple exposures to indoor contaminants: Derivation of benchmark doses and relative potency factors based on male reprotoxic effects, Regulatory Toxicology and Pharmacology 74 (2016) 23–30. [PubMed: 26644063]

[64]. Wignall JA, Shapiro AJ, Wright FA, Woodruff TJ, Chiu WA, Guyton KZ, Rusyn I, Standardizing benchmark dose calculations to improve science-based decisions in human health assessments, Environ Health Perspect 122(5) (2014) 499–505. [PubMed: 24569956]

[65]. Pham L, Watford S, Paul-Friedman K, Fostel J, Wignall J, Shapiro A, Python BMDS: A Python interface library and webserver for the canonical EPA dose-response modeling software, In preparation.

[66]. Allen BC, Kavlock RJ, Kimmel CA, Faustman EM, Dose-response assessment for developmental toxicity. II. Comparison of generic benchmark dose estimates with no observed adverse effect levels, Fundamental and applied toxicology : official journal of the Society of Toxicology 23(4) (1994) 487–95.

[67]. Filipsson AF, Sand S, Nilsson J, Victorin K, The benchmark dose method--review of available models, and recommendations for application in health risk assessment, Critical reviews in toxicology 33(5) (2003) 505–42. [PubMed: 14594105]

[68]. Yoon M, Blaauboer BJ, Clewell HJ, Quantitative in vitro to in vivo extrapolation (QIVIVE): An essential element for in vitro-based risk assessment, Toxicology 332 (2015) 1–3. [PubMed: 25680635]

[69]. Watford SM, Grashow RG, De La Rosa VY, Rudel RA, Friedman KP, Martin MT, Novel application of normalized pointwise mutual information (NPMI) to mine biomedical literature for gene sets associated with disease: use case in breast carcinogenesis, Computational Toxicology (2018).

[70]. OECD, The Global Portal to Information on Chemical Substances, eChemPortal, 2014.

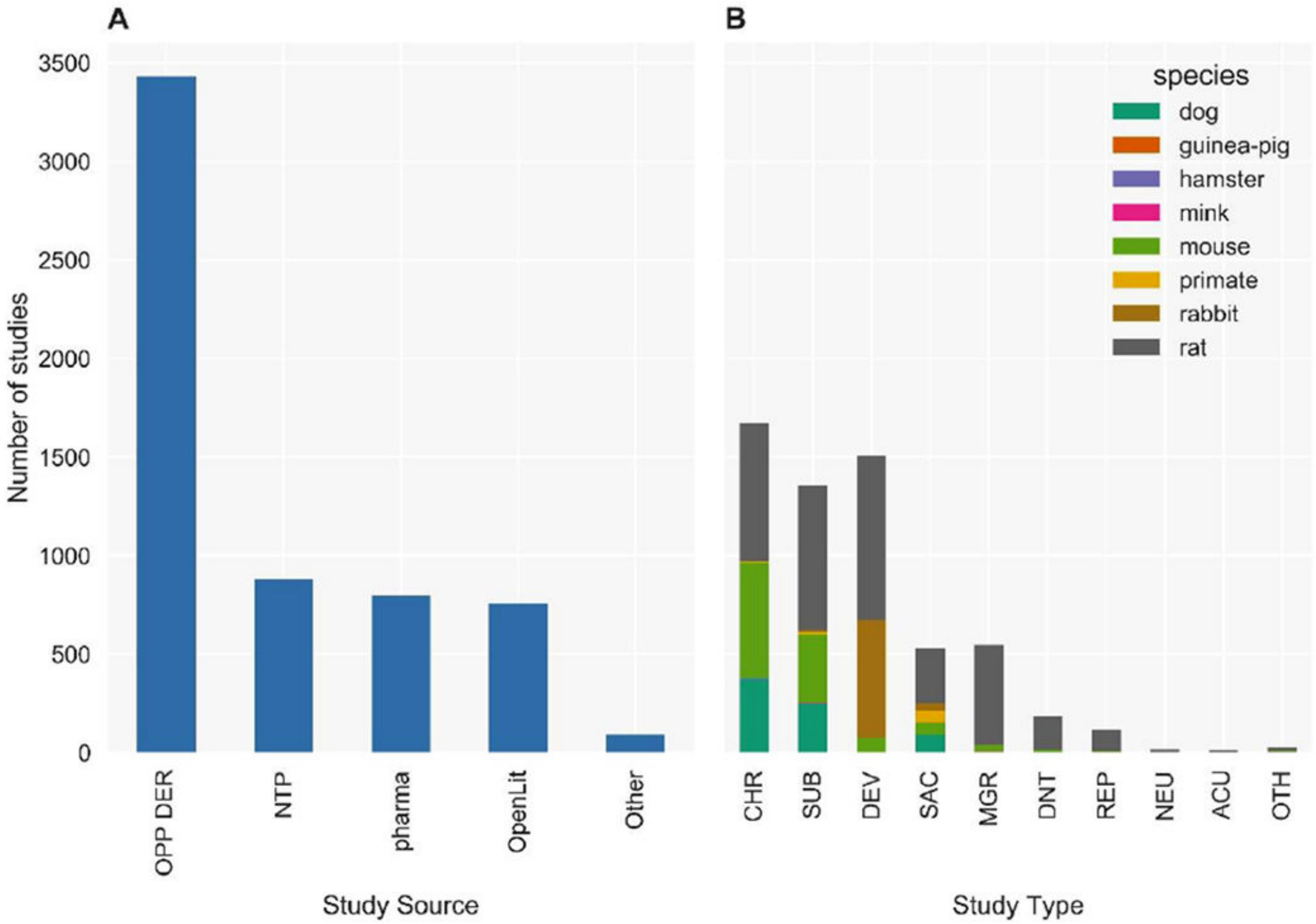


Figure 1: Number of studies by study type and species in ToxRefDB v2.

(A) ToxRefDB contains over 5,900 animal toxicity studies from a variety of sources include Office of Pesticides Programs Data Evaluation Records (OPP DER), National Toxicology Program study reports (NTP), pharmaceutical preclinical testing (pharma), open literature (OpenLit), and others (Other). (B) The study designs include chronic (CHR), sub-chronic (SUB), developmental (DEV), subacute (SAC), multigeneration reproductive (MGR), developmental neurotoxicity (DNT), reproductive (REP), neurotoxicity (NEU), acute (ACU), and other (OTH) for numerous species, but mostly for rat, mouse, rabbit, and dog.

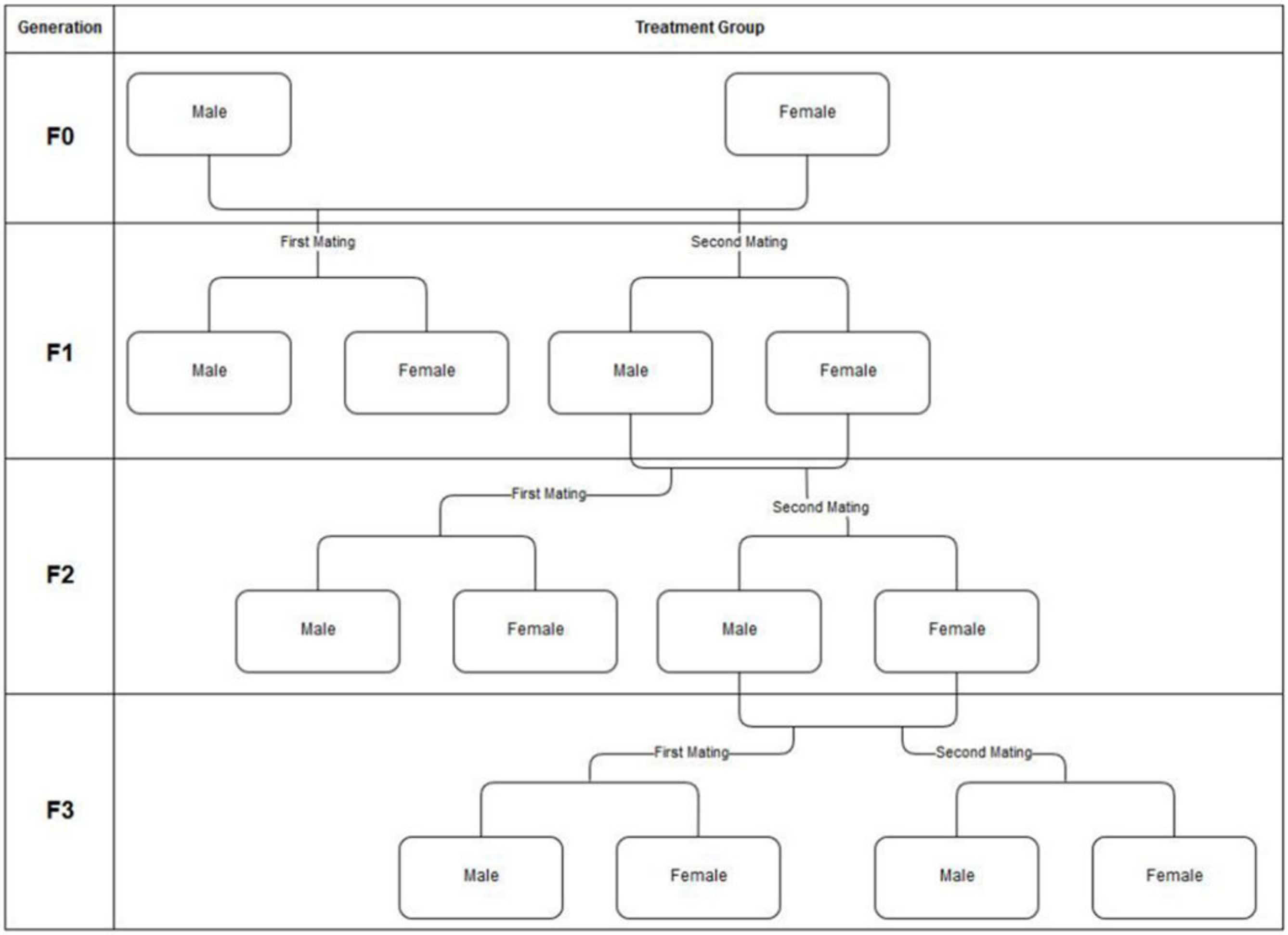


Figure 2: Three generation MGR example.

This example demonstrates that within the MGR study design, there could be 14 treatment groups, which would then need to be multiplied by the number of doses used in the study. Many of the study designs in ToxRefDB have the potential for the addition of interim, recovery, and satellite groups in order to investigate findings of interest. Even though the MGR guideline [33] does not require an F3 generation, many studies will report findings from at least a “first mating” treatment group of the F3 generation.

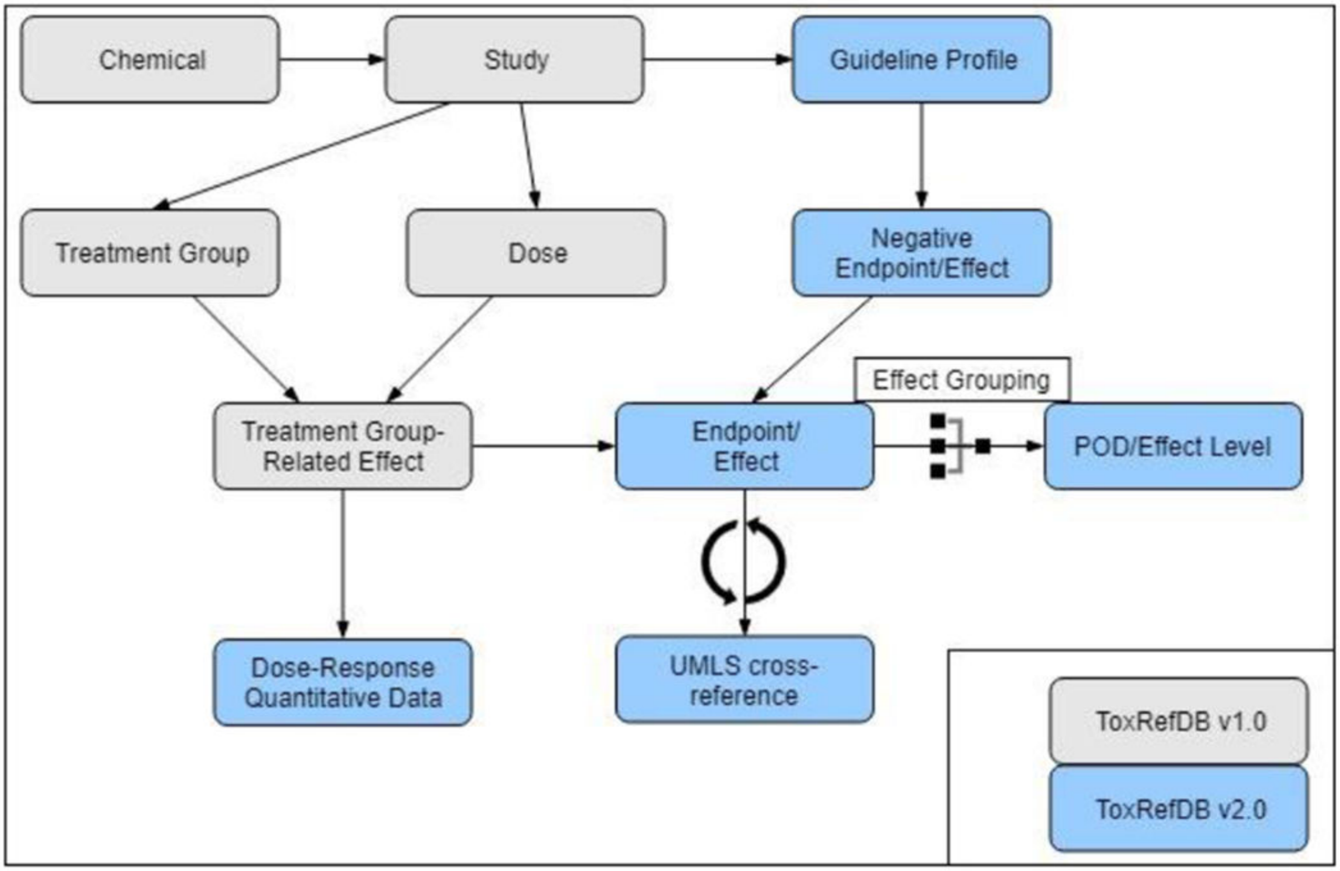
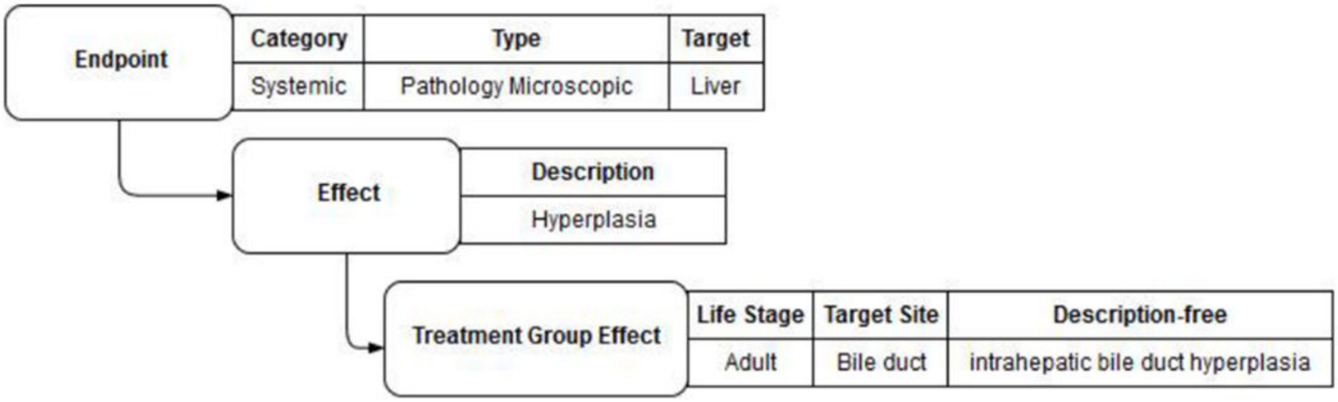


Figure 3: ToxRefDB general schema with changes made from ToxRefDB v1 to ToxRefDB v2.
Highlighted in blue are the additions to the generic schema to accommodate the updates and additional features for ToxRefDB v2. These include tables to capture the dose-response, quantitative data; guideline profiles for the inference workflow to determine negative endpoints and effects; UMLS cross-references; and effect groupings for systematically calculating PODs and associated effect levels.



Category	Type	Target
Systemic	Pathology Microscopic	Liver

Description
Hyperplasia

Life Stage	Target Site	Description-free
Adult	Bile duct	intrahepatic bile duct hyperplasia

Figure 4: Example of the controlled effect terminology in ToxRefDB v2.

An example of the terminology hierarchy is demonstrated for an effect described as “intrahepatic bile duct hyperplasia”. The finding is recorded as the “effect description free”, which is the wording used in the study report. The remaining fields are part of the ToxRefDB controlled terminology. The endpoint category is systemic, the endpoint type is pathology microscopic, the endpoint target is the liver, the effect description is hyperplasia, and the specific observation of “intrahepatic bile duct hyperplasia” was made in the adult life-stage at the specific target site, the bile duct.

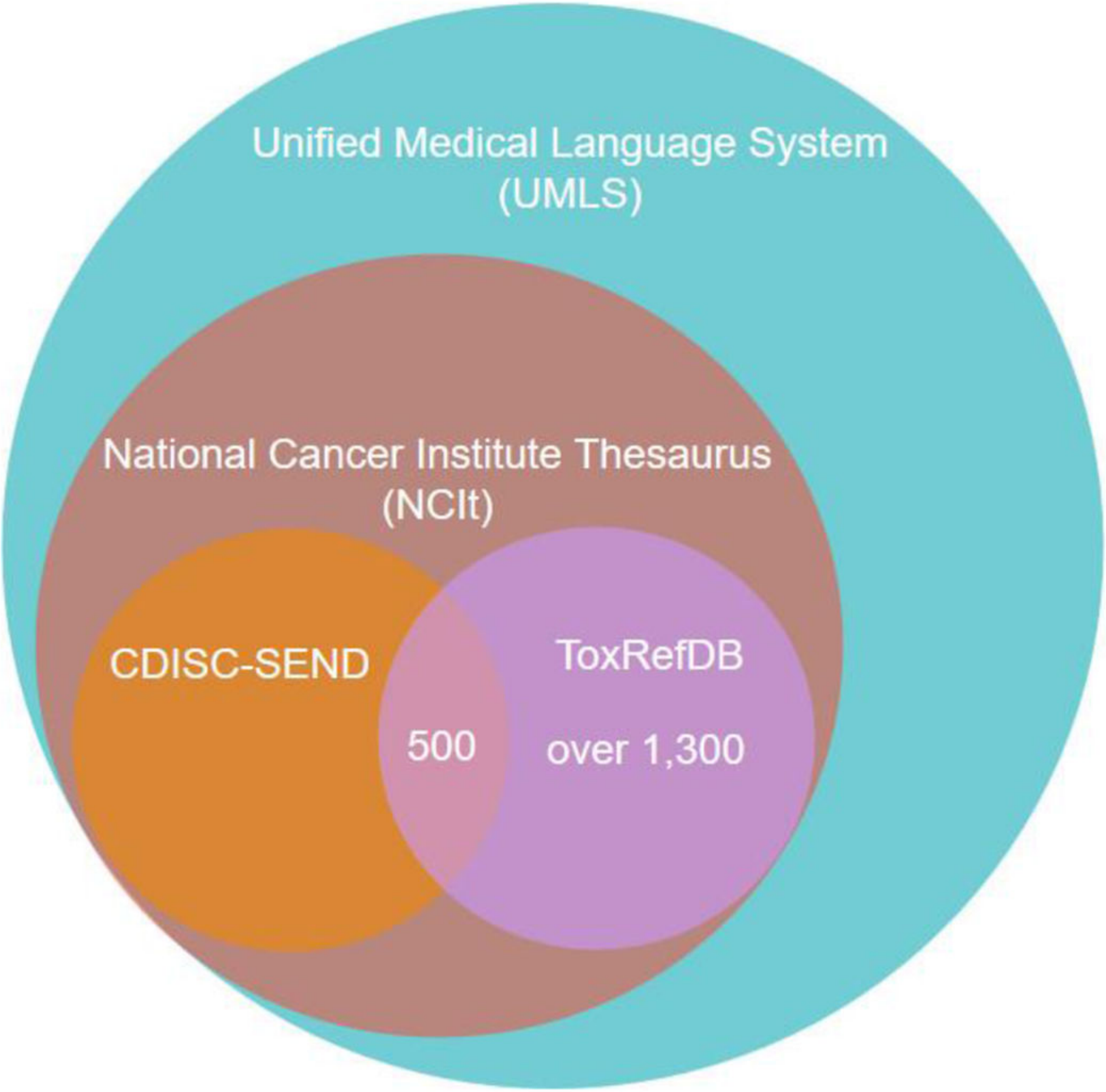

Unified Medical Language System (UMLS)
National Cancer Institute Thesaurus (NCIt)
CDISC-SEND
ToxRefDB
500
over 1,300

Figure 5: Terminology sources and membership of UMLS concept codes cross-referenced to ToxRefDB endpoints and effects.

Over 1,300 UMLS concept codes were mapped to endpoints and effects in ToxRefDB. Only 500 of those concept codes are a part of the CDISC-SEND terminology. All of the concept codes are a part of vocabularies within both National Cancer Institute Thesaurus (NCIt) as well as UMLS.

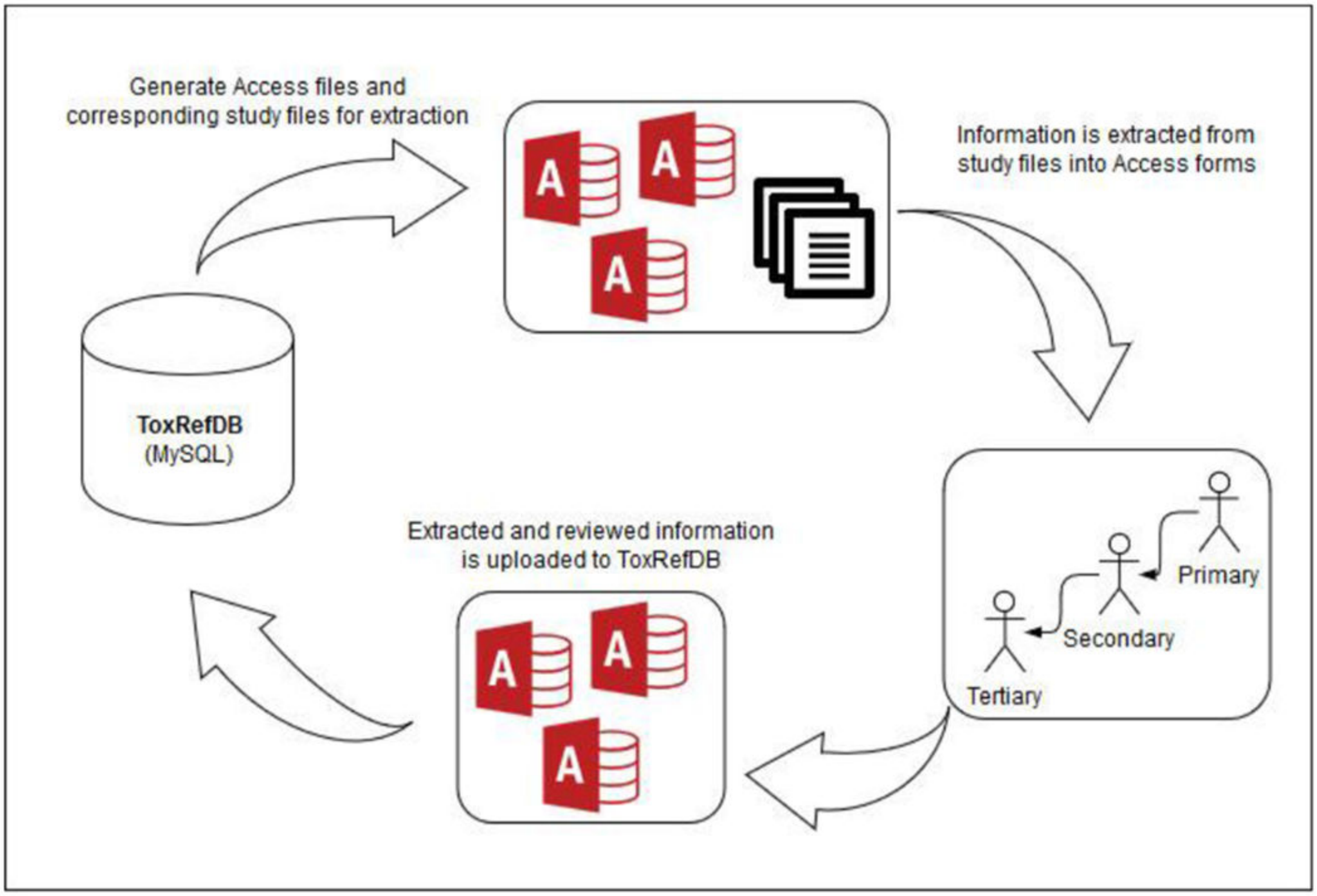

Generate Access files and corresponding study files for extraction
Information is extracted from study files into Access forms
ToxRefDB (MySQL)
Primary
Secondary
Tertiary
Extracted and reviewed information is uploaded to ToxRefDB

Figure 6: Data extraction and review workflow.

Access databases are generated for each study and bundled with the corresponding source files for data extraction. The data in the Access databases are curated with additional data extracted from the source files with up to three levels of review. The Access databases are returned by the reviewers and the data is imported back into the MySQL database.

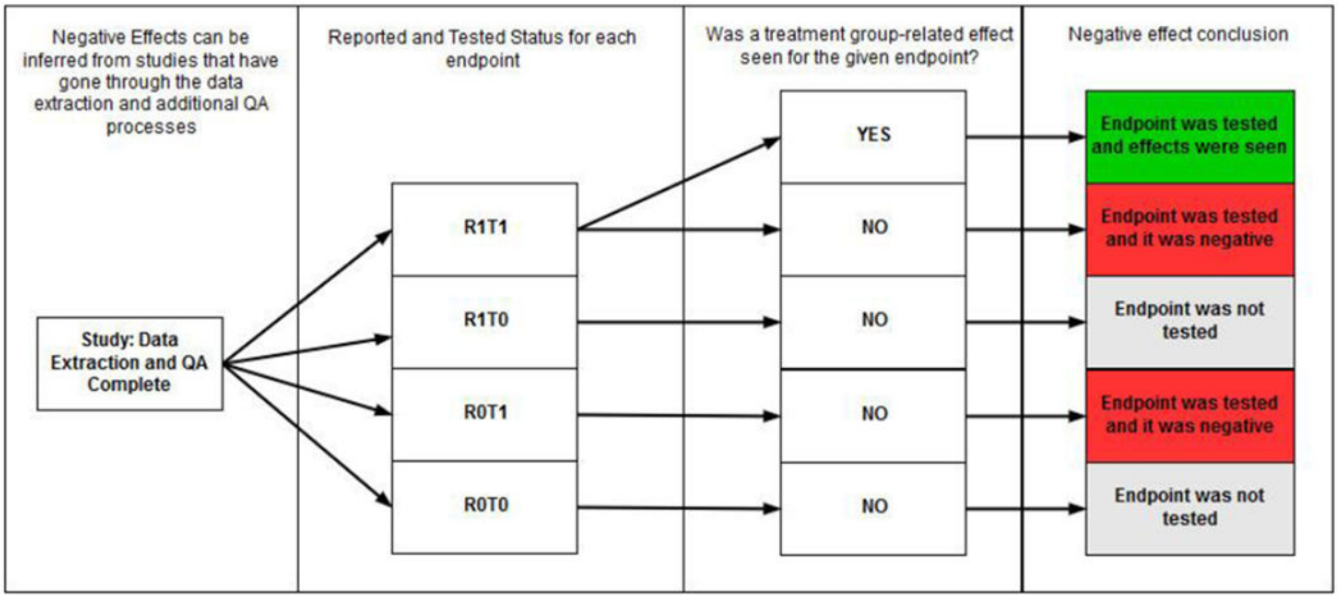

Negative Effects can be inferred from studies that have gone through the data extraction and additional QA processes
Reported and Tested Status for each endpoint
Was a treatment group-related effect seen for the given endpoint?
Negative effect conclusion
Study: Data Extraction and QA Complete
R1T1
R1T0
R0T1
R0T0
YES
NO
NO
NO
NO
Endpoint was tested and effects were seen
Endpoint was tested and it was negative
Endpoint was not tested
Endpoint was tested and it was negative
Endpoint was not tested

Figure 7: Inference workflow to determine negative effects.

Four steps are taken to systematically infer positive and negative responses: (1) study extracted completely; (2) application of the observation status; (3) determination of the effect seen (yes/no) on the basis of statistically significant findings; (4) conclusion, with positive (green), negative (red), not tested (orange), and inconclusive (gray) as possible outcomes. For Step (2), R1T1 = effect was reported and tested; R1T0 = effect was reported as not tested; R0T1 = effect was not reported but was tested; and, R0T0 = effect was not reported and not tested. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article).

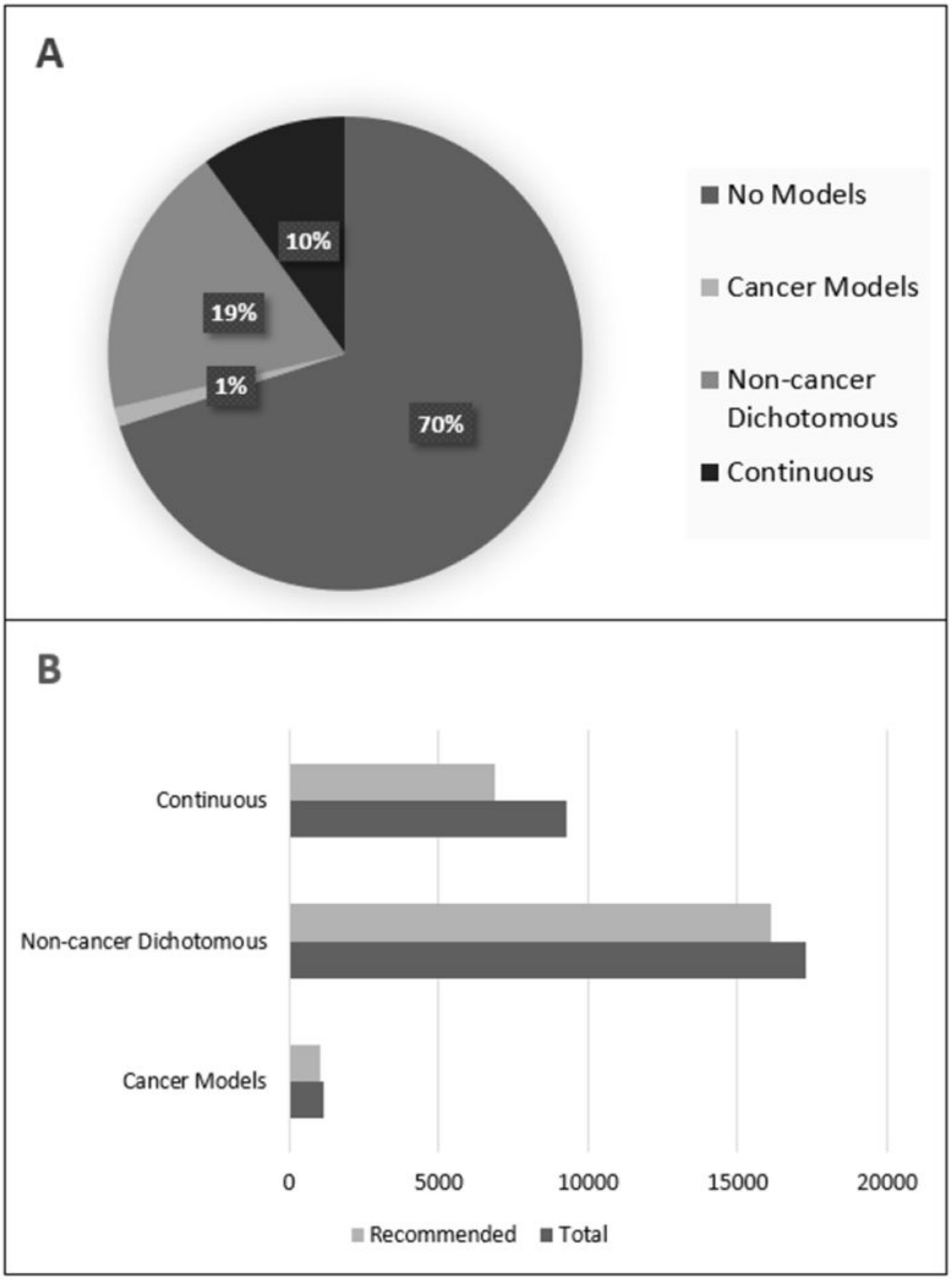


Figure 8: Proportion of effects in ToxRefDB that can be modeled.

(A) Of the 92,646 quantitative dose response datasets in ToxRefDB v2, 27,756 met data requirements for BMDS. (B) Of the datasets that met data requirements, 87% produced a recommended model result. The percentage of the data corresponding to each data type (10% BMR for dichotomous) that yielded winning models are shown.

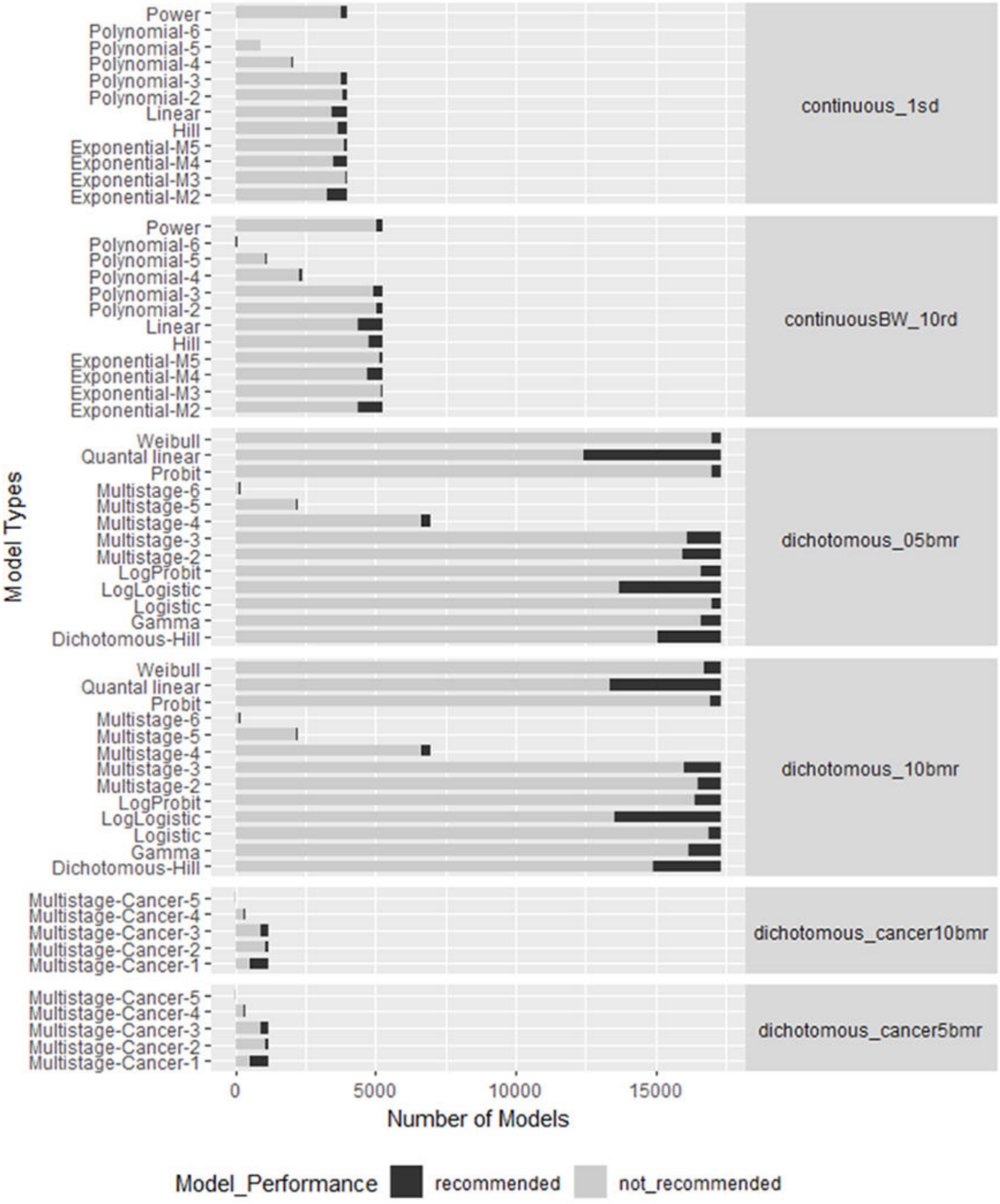


Figure 9: The number and type of models for each dataset.

The stacked bars for each model indicates the number of models that were or were not recommended. For each dataset type, the BMR was also indicated with their label. A 10 % relative deviation was used as the BMR for the continuous datasets for the body weight and organ weights. All other continuous datasets used a BMR of 1 standard deviation. The dichotomous datasets used a 5 and a 10% BMR.

Table 1:

Extraction progress as of ToxRefDB v2 release.

Over 65% of the studies have been curated with dose-response, quantitative data extracted. Priority was given to chronic (CHR) and sub-chronic (SUB) OPP DERs and NTP study reports, which are completed. The remaining studies are predominantly from the published literature and pharmaceutical companies. A) The number of studies per source by study type. B) The number of chemicals per study type.

A. Number of studies by source		
Study type	**Study source**	**Number of studies extracted**
CHR	NTP	347
CHR	OPP DER	1,079
CHR	OpenLit	9
DEV	NTP	10
DEV	OPP DER	958
DEV	OpenLit	1
DEV	Other	6
MGR	OPP DER	345
MGR	OpenLit	1
MGR	Other	20
SAC	NTP	59
SAC	OPP DER	25
SUB	NTP	247
SUB	OPP DER	769
SUB	OpenLit	6
Total		3,882

B: Number of chemicals for each study type	
Study type	**Number of chemicals**
ACU	10
CHR	663
DEV	710
DNT	124
MGR	458
NEU	18
OTH	18
REP	77
SAC	191
SUB	659
Total chemicals	1142

Table 2:

OCSPP 870 series health effects guidelines in ToxRefDB.

Guideline number	Guideline name	Study Type in ToxRefDB
870.3100	90-day Oral Toxicity in Rodents	SUB
870.3150	90-day Oral Toxicity in Nonrodents	SUB
870.3250	90-day Dermal Toxicity	SUB
870.3465	90-Day Inhalation Toxicity	SUB
870.3550	Reproduction/Development Toxicity Screening Test	REP
870.3700	Prenatal Developmental Toxicity Study	DEV
870.3800	Reproduction and Fertility Effects	MGR
870.4100	Chronic Toxicity	CHR
870.4200	Carcinogenicity	CHR
870.4300	Combined Chronic Toxicity/Carcinogenicity	CHR
870.6200	Neurotoxicity Screening Battery	NEU
870.6300	Developmental Neurotoxicity Study	DNT
870.3050	28-day Oral Toxicity in Rodents	SAC

Table 3.

Observations for guideline profiles.

The tested status indicates if the endpoint was evaluated or not by the given study. The reported status indicates if the testing status was reported in the given study. Combining the tested and reported status yields the observation status for the specific endpoint of interest on a study-by-study basis.

Tested status	**Reported status**	**Example**
Tested	Reported	The endpoint was SPECIFICALLY written in the text of the study source indicating that data was collected (default if required by the guideline for that study type)
Not tested	Reported	The endpoint was SPECIFICALLY written in the text of the study source indicating that data was NOT collected, even if required by the guideline
Tested	Not reported	The endpoint was NOT specifically written in the text of the study source, however other evidence indicates it can be deduced that it was tested (or was required by the guideline to be tested)
Not tested	Not reported	The endpoint was NOT specifically written in the text of the study source and is not required by the guideline, so we assume that the endpoint was not collected in this study

Table 4:

Guideline adherence scoring added to ToxRTool.

Score	Description
5	Adheres to modern* OECD/EPA guideline for repeat-dose toxicity studies (explicitly stated by authors; broad endpoint coverage and ability to assess dose-response)
4	Adheres to an existing or previous guideline (explicitly stated by authors; previous version of OECD/EPA guidelines or FDA guidelines)
3	Not stated to adhere to guideline but guideline-like in terms of endpoint coverage and ability to assess dose-response (e.g., NTP). Please see Quick Guide to EPA Guidelines for chronic and subchronic studies. In this table, you can easily assess whether the study was guideline-like in terms of the animals used (species, sex, age, number), dosing requirements, and reporting recommendations.
2	Unacceptable adherence to guideline (intended to adhere to guideline but had major deficiencies)
1	Unacceptable (no intention to be run as a guideline study, purely open literature or specialized study)

*A study is considered as adhering to “modern” OECD/EPA guidelines if it was published after 1998, which is the date that many Health Effect 870 series guidelines were re-published. Note that many of the studies extracted, particularly from sources like the NTP and OpenLit, were never intended to adhere to a guideline and as such “unacceptable” in this case only refers to their guideline adherence and not the study design itself.

Table 5:
Effect profiles in ToxRefDB v2 for POD computation.

Two effect profiles group effects for computation of POD values, i.e. NEL, LEL, NOAEL, and LOAEL values, which can be used in research applications.

Effect profile id description				Example output
1	Endpoints are grouped by study type, life stage, and endpoint category to produce a POD.			NEL, LEL, NOAEL, and LOAEL will be presented for combinations, e.g.: MGR/systemic/adult CHR/cholinesterase/adult DEV/systemic/fetal DEV/reproductive/adult-pregnancy
	Study type • SAC • SUB • CHR • DEV • MGR • OTH	Endpoint category • Cholinesterase • Reproductive • Developmental • Systemic	Lifestage • Adult • Adult_pregnancy • Juvenile • Fetal	
2	Endpoints are grouped according to endpoint category, endpoint type, or endpoint target. The endpoint target is used for systemic pathology endpoints, i.e. if the effects are organ-level, the POD is reported for the organ system.			NEL, LEL, NOAEL, and LOAEL will be presented for combinations, e.g.: Cholinesterase Reproductive Developmental Systemic/liver Systemic/clinical chemistry
	Endpoint category • Cholinesterase • Reproductive • Developmental			
	For the systemic endpoint category, either the endpoint type or organ are used			
	Endpoint Type: • Clinical chemistry • Hematology • In life observation For pathology microscopic and gross, and organ weights, the organ name is used, e.g.: liver, kidney, heart			

Table 6:

Number of datasets in ToxRefDB v2 for BMDS modeling.

Each dataset consists of a chemical-effect pair with all doses, number of test animals, effect values, and variance information if available. The table shows the number of modeled datasets, the BMR used, the number of recommended models, and number of chemicals with a recommended BMD model by type of data. The chemical counts are by data type, as the same chemical can have data in multiple data types. The "total" number of chemicals refers to the total number of chemicals associated with the datasets available for BMDS.

Data type	Number of datasets available for BMDS	Benchmark responses used	Have Recommended BMD Model	Chemicals (n) with a Recommended BMD Model / Total
Cancer	1 170	5%	1 101	243 / 247
		10%	1 107	246 / 247
Non-cancer dichotomous	17 318	5%	16 059	609 / 612
		10%	16 165	611 / 612
Continuous body/organ weight	9 268	10% relative deviation	4 151	416 / 430
Continuous non-body/ organ weight		1 standard deviation	3 006	284 / 300