**U.S. Environmental Protection Agency**

**Office of Research and Development**

Center for Computational Toxicology and Exposure (CCTE)

**Quality Assurance Project Plan (QAPP)**

**Title:**

“Concentration vs. Time” Toxicokinetic Chemical Concentration Time-course Database

**QA Category: B**

Authors:

John Wambaugh, Risa Sayre, and Chris Grulke

Prepared by:

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

John Wambaugh Date

Principal Investigator

Approved by:

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Garrett Nelson Date

QA Manager

# Table of Contents

[Pharmacokinetic Parameter Estimation Using R Package “invivoPKfit” 1](#_Toc25173112)

[Table of Contents 2](#_Toc25173113)

[Part A. Project Management 3](#_Toc25173114)

[A1. Description of the EPA Quality System 3](#_Toc25173115)

[A2. Distribution List 4](#_Toc25173116)

[A3. Project/Task Organization and Responsibilities 5](#_Toc25173117)

[A4. Problem Background 5](#_Toc25173118)

[A5. Task Description and Schedule 7](#_Toc25173119)

[Product 1 7](#_Toc25173120)

[Product 2 7](#_Toc25173121)

[Product 3 7](#_Toc25173122)

[Product 4 7](#_Toc25173123)

[Product 5 7](#_Toc25173124)

[Product 6 7](#_Toc25173125)

[A6. Quality Objectives and Criteria for Model Inputs/Outputs 7](#_Toc25173126)

[A7. Specialized Training Requirements/Certification 8](#_Toc25173127)

[A8. Documentation and Records 8](#_Toc25173128)

[Part B. Assessment and Oversight 9](#_Toc25173129)

[B1. Assessment and Response Actions 9](#_Toc25173130)

[B2. Reports to Management 9](#_Toc25173131)

[Part C. Estimation Method Development 9](#_Toc25173132)

[C1. Parameter Estimation Methodology 9](#_Toc25173133)

[C2. PK/TK Models 10](#_Toc25173134)

[C3. Model Selection 12](#_Toc25173135)

[C4. Error Propagation with Quadrature 12](#_Toc25173136)

[C5. Statistical Significance Testing 12](#_Toc25173137)

[C6. Random Number Seeds 13](#_Toc25173138)

[C7. Review, Verification, and Validation 13](#_Toc25173139)

[Part D. R Package Maintenance 13](#_Toc25173140)

[D1. Method Coding 13](#_Toc25173141)

[D2. Using the R “tests” Directory Functionality 14](#_Toc25173142)

[D3. Uploading to GitHub 14](#_Toc25173143)

[Part E. Manuscript Documentation 15](#_Toc25173144)

[Part F. References 15](#_Toc25173145)

# Part A. Project Management

## A1. Description of the EPA Quality System

Disclaimer

The EPA does not consider this internal planning document an official Agency dissemination of information under the Agency’s Information Quality Guidelines, because it is not being used to formulate or support a regulation or guidance; or to represent a final Agency decision or position. This planning document describes the overall quality assurance approach that will be used during the research study. Mention of trade names or commercial products in this planning document does not constitute endorsement or recommendation for use.

The EPA Quality System

The EPA requires that all data collected for the characterization of environmental processes and conditions are of the appropriate type and quality for their intended use. This is accomplished through an Agency-wide quality system for environmental data. Components of the EPA quality system can be found at <http://www.epa.gov/quality>. EPA policy is based on the national consensus standard ANSI/ASQ E4-2004 Quality Systems for Environmental Data and Technology Programs: Requirements with Guidance for Use.

This QAPP provides information concerning refinement, evaluation, and support of a suite of models and data for predicting the toxicokinetics of chemicals using *in vitro* data as well as *in vitro-in vivo* extrapolation of high throughput screening data.

## A2. Distribution List

|  |  |  |  |
| --- | --- | --- | --- |
| **Name** | **Affiliation** | **Title/Role** | **Email** |
| Wambaugh, John | CCTE | Technical Lead | [wambaugh.john@epa.gov](mailto:wambaugh.john@epa.gov) |
| Nelson, Garrett | CCTE | Quality Assurance Manager (QAM) | [nelson.garrett@epa.gov](mailto:nelson.garrett@epa.gov) |
| ***Vacant*** | CCTE | Database Curator and Programmer |  |
| Katie Paul-Friedman | CCTE | Principal Investigator / CSS ISI Output Lead | [paul-friedman.katie@epa.gov](mailto:paul-friedman.katie@epa.gov) |
| Sayre, Risa | CCTE | Ph.D. Student / Database Designer | [sayre.risa@epa.gov](mailto:sayre.risa@epa.gov) |
| Grulke, Christopher | CCTE | Principal Investigator / Database Designer | grulke.christopher@epa.gov |
| Williams, Antony | CCTE | Principal Investigator / CompTox Chemicals Dashboard Architect | [williams.antony@epa.gov](mailto:lau.christopher@epa.gov) |
| Evans, Marina | CCTE | Principal Investigator / Toxicokinetic Modeling | [evans.Marina@epa.gov](mailto:evans.Marina@epa.gov) |
| Kenyon, Elaina | CCTE | Principal Investigator / Toxicokinetic Modeling | [kenyon.Elaina@epa.gov](mailto:kenyon.Elaina@epa.gov) |
| El-Masri, Hisham | CCTE | Principal Investigator / Toxicokinetic Modeling | [el-Masri.Hisham@epa.gov](mailto:el-Masri.Hisham@epa.gov) |
| Tornero-Velez, Rogelio | CCTE | Principal Investigator / Toxicokinetic Modeling | [tornero-velez.rogelio@epa.gov](mailto:tornero-velez.rogelio@epa.gov) |
| Kapraun, Dustin | CPHEA | Principal Investigator / Toxicokinetic Modeling | [kapraun.dustin@epa.gov](mailto:kapraun.dustin@epa.gov) |
| Schlosser, Paul | CPHEA | Principal Investigator / Toxicokinetic Modeling | [schlosser.Paul@epa.gov](mailto:schlosser.Paul@epa.gov) |
| DeVito, Michael | CCTE | Division Director / Chemical Characterization and Exposure Division | [devito.Michael@epa.gov](mailto:devito.Michael@epa.gov) |
| Egeghy, Peter | CCTE | Branch Chief / Computational Exposure and Toxicokinetics | [egeghy.peter@epa.gov](mailto:egeghy.peter@epa.gov) |
| Hughes, Michael | CCTE | Branch Chief / Experimental Toxicokinetics and Toxicodynamics | [hughes.MichaelF@epa.gov](mailto:hughes.MichaelF@epa.gov) |
| Dunne, Jeremy | CCTE | Branch Chief / Application Development | [dunne.jeremy@epa.gov](mailto:dunne.jeremy@epa.gov) |

## A3. Project/Task Organization and Responsibilities

This Quality Assurance Project Plan (QAPP) was developed with reference to Guidance for Quality Assurance Project Plans for Modeling EPA QA/G-5M (U.S. Environmental Protection Agency, 2002).

This QAPP covers a project that supports R package “invivoPKfit”. This open source package allows statistical estimation of pharmacokinetic parameters describing chemical absorption, distribution, metabolism, and excretion by the body. The following list provides the project participants along with their respective responsibilities.

John Wambaugh is a principal investigator in the CCTE. He is responsible for supervising 1) the population of data into the toxicokinetic concentration vs. time (***CvT***) database, 2) the development of interfaces (such as the dashboard) to the data, and 3) any revisions to the structure of the database. He is also responsible for overseeing the development and writing of technical manuscripts.

Garrett Nelson is the Quality Assurance Manager (QAM) in the CCTE. He will receive copies of all QA and QC related information. He will review QAPP and any pertinent QA related documents to ensure the quality of the research is upheld. He will perform, at his discretion, at least one external Technical Systems Assessment (TSA) during the project

A currently vacant position within CCTE will serve two primary roles for the ***CvT*** database. This individual will first populate the database with data by scouring the toxicokinetic literature and interacting with researchers who have data sets to contribute. This individual’s second role will be to develop interfaces to the database, including the CompTox Chemical Dashboard.

Risa Sayre is an ORISE doctoral student in the Department of Environmental Science and Engineering at the University of North Carolina. She maintains an office and an active role in the CCTE, and her research supports both EPA and her dissertation. Risa and Chris Grulke created the toxicokinetic concentration vs. time (***CvT***) database and has used the invivoPKfit software to estimate parameters from that database.

Chris Grulke is a principal investigator in the CCTE. Chris and Risa Sayre created the toxicokinetic ***CvT*** database.

Marina Evans, Elaina Kenyon, Hisham El-Masri, and Rogelio Tornero-Velez are toxicokinetic mathematical modeling principal investigators working in CCTE, while Dustin Kapraun and Paul Schlosser are toxicokinetic mathematical modeling principal investigators working in CPHEA. Their responsibilities include providing TK data sets for archiving within the ***CvT*** database and drawing upon the database for evaluation of models.

Michael Devito, Peter Egeghy, and Michael Hughes are managers within CCTE who supervise personnel who may regularly contribute to and use the ***CvT*** database.

Antony Williams is a principal investigator in CCTE and is architect of the public-facing CompTox Chemicals Dashboard (<https://comptox.epa.gov/>). When provided with updated chemical- or product-specific exposure measurements (for example, chemical concentrations observed in household products), he works to provide those predictions to the public.

Katie Paul-Friedman is a principal investigator in CCTE who is leads “Data Compilation and Curation for Risk Assessment” within the Chemical Safety for Sustainability (CSS) research program. She oversees continued development and curation of databases to support chemical safety decision making, including mammalian toxicity, exposure, and new approach methodology (NAM) data.

Jeremy Dunne is a Branch Chief within CCTE who supervise personnel who may regularly contribute to the integration of the ***CvT*** database with public-facing resources including the CompTox Chemicals Dashboard.

## A4. Problem Background

The study of toxicokinetics (TK) and/or pharmacokinetics (PK) allows understanding of the body’s absorption, distribution, metabolism, and excretion of chemicals or pharmaceuticals, respectively. TK is a critical component in understanding toxicological responses by linking chemical exposure to internal tissue concentrations. TK/PK studies typical consist of controlled dosing of live subjects to a chemical, followed by serial measurement of chemical concentration in different matrices (for example, serum or urine). Sacrificial animal studies may kill, dissect, and analyze specific tissues at sequential times. These *concentration time-course data* are referred to variously as “concentration vs. time” or ***“CvT”*** data. TK/PK involves mathematically describing TK/PK data such that extrapolation is possible. Time courses of compound concentrations in plasma are used in chemical safety analysis to evaluate the relationship between external administered doses and internal tissue exposures.

The ***CvT*** database is intended to serve as a comprehensive repository for TK/PK data, including published sources of ***CvT*** data, ***CvT*** concentration vs time series, and derived pharmacokinetic parameters. While many TK experiments have been published, there has been little effort to normalize, structure, and centralize the results of these studies. An international workshop held in February 2016 focused on key steps needed to facilitate the adoption of high throughput TK into chemical risk prioritization and decision making(S. M. Bell et al., 2017). That workshop recommended the “Creation of a database that could house all shared *in vitro* and *in vivo* TK data, and identification of actions to be taken to encourage sharing of existing data.” (S. M. Bell et al., 2017)

Here, we present a public database for storing ***CvT*** data and its associated study metadata. The compiled ***CvT*** data have been analyzed with TK curve-fitting software to add a set of uniformly estimated properties, such as area under the concentration-time curve (AUC), volume of distribution (Vd), and elimination half-life, to the database. It is hoped that these data will serve as a catalyst for the public sharing of curated TK data to improve assessment of risk posed by chemicals to human health.

The current dataset contains a large subset of data based on an initial, externally-identified list of chemicals, as well as the easily obtainable already-structured datasets. However, it covers only a small proportion of the extent of the tested chemical space and doesn’t necessarily contain all instances of concentration time-series data for any of the chemicals therein. All major administration routes are incorporated within the database (Figure 1), with concentrations measured in blood/plasma, tissues, and excreta. In addition to pharmacokinetic model calibration and validation, these data may be used for analyses of differential chemical distribution across chemicals, species, doses, or routes, and for meta-analyses on pharmacokinetic studies.

There is increasing acceptance of the use of *in vitro*-derived TK for chemical risk prioritization (Shannon M. Bell et al., 2017). However, there is still need for careful model evaluation to determine which chemicals these techniques might work for and how well (Oreskes, 1998; Wambaugh et al., 2015). The ***CvT*** database provides an excellent source to evaluate modelled relationships between external and internal doses. It is hoped that these data will serve as a catalyst for the public sharing of curated TK data to improve assessment of risk posed by chemicals to human health.

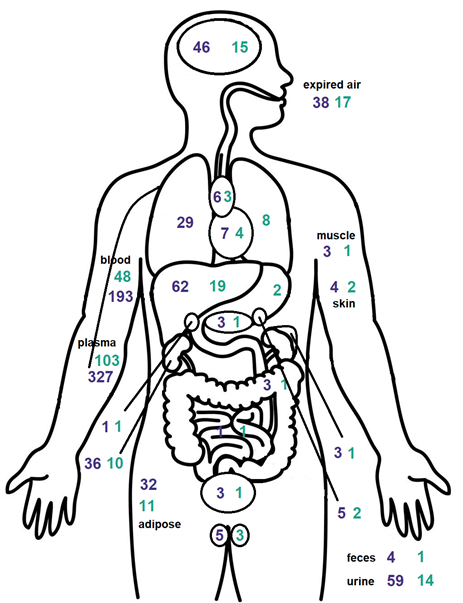


Figure : From (Sayre, Wambaugh, & Grulke, 2019), the count of studies (in purple) and test substances (in green) with **CvT** results in different media (for any species but represented on a human body).

## A5. Task Description and Approximate Schedule

The objective of this project is to populate the ***CvT*** database with additional chemical data as well as refine the database such that it more completely allows statistical evaluation of the data contained within it. A secondary objective requiring significant effort is to make the data publicly available. The data curation process is documented in a standard operating procedure (SOP) “CvTdb Data Curation” which is available from GitHub: <https://github.com/USEPA/CompTox-PK-CvTdb>

Contributing products are listed below:

### Product 1

Initial development of EPA’s ***CvT*** Database. (**FY19**)

### Product 2

Development of CompTox Chemical Dashboard interface for ***CvT*** database. (**FY20)**

### Product 3

Curation of dermal absorption route data into ***CvT*** database. (**FY20)**

### Product 4

Curation of aerosol inhalation route data into ***CvT*** database. (**FY20)**

### Product 5

Development of manuscript describing observations of the ***CvT*** database team on current reproducibility of PBTK models with recommendations for improvement. **(FY20)**

### Product 6

Curation of literature and ORD toxicokinetic datasets into ***CvT*** database. **(FY21)**

### Product 7

Classification of all chemicals in the EPA’s ***CvT*** Database with respect to restrictive- or non-restrictive metabolic clearance using information theoretic measures of parsimony. (**FY21)**

### Product 8

Estimation of tissue partition coefficients for a generic PBTK model for all chemicals in the EPA’s ***CvT*** Database. (**FY22**)

## A6. Quality Objectives and Criteria for Model Inputs/Outputs

The data curation process is documented in a standard operating procedure (SOP) “CvTdb Data Curation” which is available from GitHub: <https://github.com/USEPA/CompTox-PK-CvTdb>

The data entry template (<https://github.com/USEPA/CompTox-PK-CvTdb/raw/master/CvT_data_template_articles.xlsx>) in a Microsoft Excel workbook with five tabs. The “Conc. Time values” tab contains the concentration vs. time data, but the other tabs help explain the experiments that are being recorded. Each data point must be associated with a “Series” described int the series tab.

All information related to the project will be stored in a directory named:

**L:\Lab\NCCT\_ExpoCast\ExpoCast[YEAR]\CvT\**

The underlying data will be stored in standard machine-readable format (typically a comma separated value, CSV, file) that is easily read by different programming languages and can be easily disseminated to others via required EPA systems (for example, ScienceHub).

The database will be stored in MySql format in a directory:

**L:\Lab\NCCT\_ExpoCast\ExpoCast[YEAR]\MySQL\**

## A7. Specialized Training Requirements / Certification

Specialized expertise and qualifications needed to participate in this project include proficiency with scientific computing in modern computer programming languages and database management. Team members must be knowledgeable in at least one of the following fields: computer science, databases, biomedical engineering, mathematical physiology, applied mathematics, applied statistics, physics, mathematics, or chemistry. Team members should be fluent programmer in a modern database language (for example, MySQL) and programming language (for example, R, Python, Julia). This research project involves data science, mathematical modeling, toxicokinetics, and analysis.

A currently vacant position within CCTE will serve two primary roles for the ***CvT*** database. This individual will first populate the database with data by scouring the toxicokinetic literature and interacting with researchers who have data sets to contribute. This individual’s second role will be to develop interfaces to the database, including the CompTox Chemical Dashboard.

For trainees working on the project, a basic knowledge of programming, chemistry, biology and mathematics will be required. Additional training will be provided, if needed. Reference to this training will be documented in an electronic lab notebook that is easily accessible to all with the initials of the trainer and trainee and date of completion. If expert workshops or meetings are required to advance the working knowledge of individuals involved in the projects, the name of these workshops, the attendees, and their dates of attendance will also be placed into the electronic lab notebook. If certificates were issued for these workshops, they will be placed into a separate binder containing relevant lab documents, or, if digital, will be placed into a folder labeled “Certificates of Training” that will be located on the Principal Investigator’s computer desktop.

# Part B. Assessment and Oversight

## B1. Assessment and Response Actions

At the discretion of the researchers as well as the QAM a mid-project technical system audit may be requested by the project QAM to assess the quality assurance process of the project. Any findings, research best practices, or improvements will be compiled in a report and filed with other QA documents. Any corrective actions and responses will be documented and filed with the Assessment Report. These findings and corrective measures will be in the office of the Principal Investigator.

## A8. Documentation and Records

When a new individual becomes involved in the project, this QAPP will be sent to her or him by John Wambaugh. The individual will be encouraged to ask questions should any part of this document be unclear, and upon reading their name(s) and dates of completion will be entered into an open lab notebook or electronic record of training. Each project contributor should be given access to L:/Lab/NCCT\_ExpoCast. Each project contributor should be given a Confluence account.

## B2. Reports to Management

The director and the QAM will be apprised of progress on a semi-annual basis and at the end of the study, or as and when the data will be presented during work-in-progress meetings or at scientific meetings. The types of progress reports may include brief (i.e., 1-5 page) progress reports, presentations, scientific posters, meeting abstracts, and/or manuscript drafts. John Wambaugh will be responsible for preparing and distributing the reports.

# Part C. Database Maintenance and Revision

The data is maintained by U.S. EPA’s Center for Computational Toxicology and Exposure in the data model depicted in Figure 2 instantiated in a MySQL 5.6 community edition relational database of simple data types (text, numeric, and Boolean).

***CvT*** time-series consist of the measured *in vivo* concentration-time data points (generally extracted from figures in the paper) resulting from a toxicokinetic study and the experimental details that provide the context for that toxicokinetic curve. Each data point is stored in the **conc\_time\_values** table as seen in Figure 2. The **series** table contains details regarding each set of values, and **studies** contains more general information about the pharmacokinetic experiment. Each study is linked to **documents**, which cites the information source.

## C1. Database Structure

The original inspiration for the set of contextualizing metadata was based on the test guideline for metabolism and pharmacokinetics released by the U.S. EPA Office of Prevention, Pesticides, and Toxic Substances (OPPTS) (U.S. Environmental Protection Agency, 1998). Identifying the set of parameters necessary to properly annotate the extracted ***CvT*** data was iterative, with improvements evolving after the review of multiple publications (as the consistency and reporting of study details is highly variable in literature). Below is the set of key study details collected:

* Reference (data source document identification, the name of the figure or table from which the data was gleaned, explanatory notes provided in the source publication)
* Study scenario (test substance, administration route, dose amount, vehicle, and volume, exposure duration, quantity of doses given and their spacing, number of subjects per treatment group, number of treatment groups, fasting status of subjects)
* Subject details (species, type/strain, sex, age, age category, size, and any other description given by source)
* Measurement details (the original time and concentration units, analyte, the medium (tissue, circulatory fluid, etc.) in which the analyte was detected)
* Measurement methods (the limits of detection and/or quantification, the analysis method)
* Curation notes (known assumptions made during the collection process)

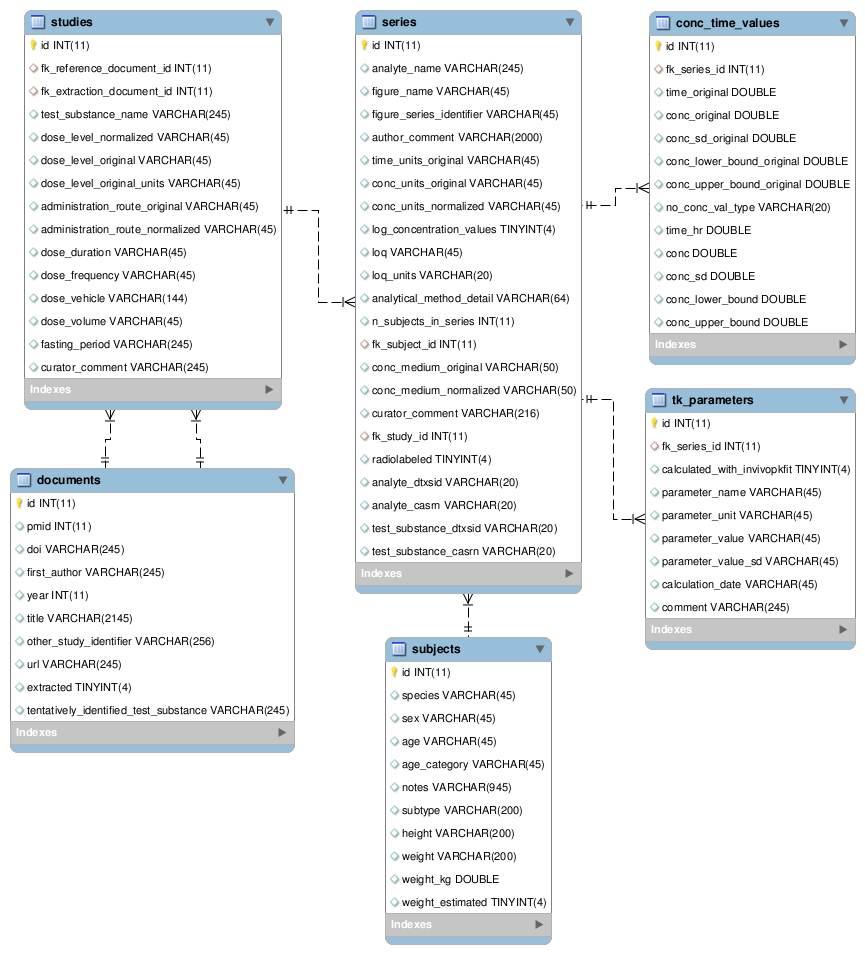


Figure : Entity-relationship diagram of CvTdb.

During extraction of the data from concentration-time plots, values indicating the extrema of any y-axis (concentration) error bars were captured for some series (collection of error bars started partly through our curation process). These were assumed to indicate the standard deviation. When the standard deviation was calculated in log space, the non-logged values were stored as conc\_lower\_bound and conc\_upper\_bound. All data was stored with the intent to maintain the original content with as little alteration as possible in the “\*\_original” fields and then normalized into corresponding “\*\_normalized” fields (for example, see the *conc\_units\_original* and *conc\_units\_normalized* fields in the **series** table).

## C2. Toxicokinetic Parameters

***CvT*** time series provided a basis for estimating several TK parameters. These parameters are stored in the **tk\_parameters** table, linked to the time series from which they were derived. A binary flag field “calculated\_with\_invivopkfit” can be used to separate parameters calculated using the standard ***CvT*** analysis package invivopkfit and the data collected from the source documents. Parameters include Cmax (the maximum concentration), total clearance, Michaelis-Menten values Km and Vmax, and elimination half-life. Distributions of the generated TK parameters can be seen in the supplementary zipfile (tk\_params.pdf).

## C3. Review, Verification, and Validation

New releases of the database are considered to be fully reviewed, verified, and validated upon EPA clearance of a peer-reviewed manuscript in which the methods are included (potentially as supplemental material). Along with the manuscript, a new version will be released to Github). The database will also be archived with EPA’s Science Inventory.

## C4. Distribution

New versions of the database are cleared as supplemental material for peer-reviewed scientific manuscripts. Once cleared, the database is made public by the CompTox Chemical Dashboard. The collected data have been made publicly available on FigShare as a zipfile containing a sql dump of the database, extracted csv files, and additional supplements:

<https://doi.org/10.23645/epacomptox.9925151>

The data are also available at the GitHub repository:

[https://github.com/USEPA/CompTox-PK-CvTdb](https://github.com/USEPA/CompTox-PK-Cvtdb)

along with a xlsx template for submission of new studies to the database and the latest versions of the supplements to this paper.

In addition, display of the data through the U.S. EPA’s Chemistry Dashboard (<https://comptox.epa.gov/dashboard> (Williams et al., 2017)) is planned but not yet implemented. The detailed ***CvT*** data is only available through the sql dump file provided in the zipfile on FigShare. A medium term goal is to make the database available through the ADME tab of the CompTox Chemicals Dashboard.

# Part D. Populating the Database

Database template location…

## D1. Documenting Data Sources

Chemical identity is critical for the proper storage and analysis of TK data. The chemical tested or analyzed is often only identified in source documents. This leaves some room for ambiguity (Richard, 2004). The source names were mapped to unique substances (designated by DSSTox SIDs) through expert curation. Identities should be established using ‘preferred\_names’ and DTXSID’s obtained from searching DSSTox (EPA’s Distributed Structure-Searchable Toxicity Database) (Grulke, Williams, Thillanadarajah, & Richard, 2019). If information cannot be found, a request to “register” a compound should be made to Chris Grulke ([grulke.chris@epa.gov](mailto:grulke.chris@epa.gov)).

### D1-1. Structured Source Data

Structured data is formatted into our data storage model shown in Figure 2. Copies of the original structured data should be stored in:

**L:\Lab\NCCT\_ExpoCast\ExpoCast[YEAR]\CvT\Sources\Structured\**

An archival copy of the database as downloaded should be stored in:

**..\CvT\Sources\Structured\Raw\**

Any processing scripts and manual interventions should be stored in:

**..\CvT\Sources\Structured\Scripts\**

The final, ready for CvT data should be placed in:

**..\CvT\Sources\Structured\Processed\**

### D1-2. Literature Source Data Extraction

Access to articles containing time-series data may be provided by the EPA library, or by publishers posting works openly online in Adobe PDF format. In some cases, publications may make tabular concentration vs time data available. For most cases, however, the data are only available in plots. These plots should be converted to images, for example by using the native screenshot tool in Windows.

Original articles, tables (if available) and extracted images should be storied in:

**L:\Lab\NCCT\_ExpoCast\ExpoCast[YEAR]\CvT\Sources\Extracted\Raw**

Images should then be processed using a data digitizer tool such as WebPlotDigitizer (<https://automeris.io/WebPlotDigitizer/>) (Rohatgi, 2018). Such tools require manual identification of axes and scales for calibration and the selection of each point for extraction by a user. Automated approaches for this task are available, but were found to be error prone in calibration, identification of units and scale, or incapable of extracting data from plots with logarithmic axes. Extracted tables of concentration vs time data can be exported to CSV (comma separated value files), then imported into the database using MySQL Workbench. The name and version of the data extraction tool should be included in the name of the CSV file.

Collecting the details of the study from which the concentration vs time data resulted is completed manually. Both annotation information and CSV files should be stored in:

**..\CvT\Sources\Extracted\Processed\**

### D1-3. Data Validity

In assessing the validity of the data collected, three aspects are considered: limitations in the source material, unreported experimental variability from ***CvT*** experiments, and inaccuracy in the collection techniques applied when extracting the data from source content.

## D2. Assessing Completeness of Records

It is important to note that the experimental details provided with the source data come with varying degrees of completeness. It is common for study details that could be relevant to interpretation and comparison between experiments to be omitted in the source material. For example, the method used to quantify chemicals in media, the recovery amount associated with that method, and its associated limit of quantification were rarely reported in our sources. To give another example, 145 of 263 oral dosing studies did not report whether subjects were fasted. Absence of study details from the database should be considered by the user when accessing the validity of the data for their use cases.

## D3. Time Series Data Normalization

The data are extracted with an effort to be as faithful to the original as possible. That means that everything was collected as reported, with no consideration for standardizing the units nor controlling the vocabulary used to describe the experiment. Standardized values were calculated and stored alongside the original values to facilitate easier comparison across experiments. These normalized concentrations are stored in µg/mL for tissues, excreta, or plasma (µgEq/mL for radiographic measurements), and µg/m^3 for breath. Times are stored in hours.

A maximum of 5 numbers after the decimal point are stored for these data types. Doses are stored in mg/kg. In cases where the mass of the subject was not reported, and the dose was administered as a simple mass and not a body weight proportion, a mg/kg dose was calculated using the average mass for all subjects of that type.

The names for the tissues in which doses were measured are standardized to the preferred names in NCI thesaurus (Sioutos et al., 2007). All subject measurements were stored in cm for height and kg for weight/mass. The subject age may be reported as a numerical value, a category, or both.

## D4. Toxicokinetic Parameter Calculation

Toxicokinetic parameters are often used as surrogates for more complicated time-series data when prioritizing chemicals according to their TK profiles; however, the models used to derive these parameters can be inconsistent between different publications, making direct comparisons problematic. Hence, a systematic extraction of reported TK parameters from literature was not done as part of this databasing effort (though several TK parameter estimations reported in the literature were gathered). Rather, the R package invivoPKfit (Wambaugh, Ring, & Setzer, 2019) is used to calculate pharmacokinetic parameters using the function fit\_all as described in previous publications (Wambaugh et al., 2018).

An aggregate set of these parameters were computed for all plasma concentration data for each administration route:test substance pair, as described in (Wambaugh et al., 2018). The optimized likelihood for both the one- and two-compartment models were compared using the Akaike Information Criterion (AIC) (Akaike, 1974). Study-specific standard deviations were included in the number of parameters used to calculate AIC (for example, if there were data from two studies, there were two standard deviation parameters estimated and factored into the AIC). The results of the model with the lesser AIC (Burnham, 2003) are stored in the database. The script used to calculate the parameters is publicly available through GitHub repository:

<https://github.com/USEPA/CompTox-ExpoCast-invivoPKfit>

# Part E. Manuscript Documentation

Each manuscript should have an archival directory named:

L:/ExpoCast/ExpoCast[YEAR]/[PROJECTNAME]

A final working data processing script should be stored along with the raw data on the L:/ drive in a directory called:

L:/ExpoCast/ExpoCast[YEAR]/[PROJECTNAME]/rawdata/

This data processing code should take raw data and convert it to the form used by the paper to make figures. A final working script to make all figures associated with the paper and any necessary processed data should be stored in

L:/ExpoCast/ExpoCast[YEAR]/[PROJECTNAME]/figures/

Multiple copies of manuscript milestones should be stored:

|  |  |
| --- | --- |
| Milestone | Sub-Directory |
| Initial drafts, figures, scripts, and other material | EarlyDrafts |
| Version submitted to tech review and edits | TechReview |
| Version submitted to clearance review and edits | ClearanceReview |
| Version submitted to journal | Submission |
| Any revisions submitted to journal and edits | Revision[#] |
| Any submissions to new journal | Submission[#] |

Each manuscript should be accompanied by a new vignette (<http://r-pkgs.had.co.nz/vignettes.html>) that demonstrates how to create the appropriate figures from the paper using the new version of the httk package.

# Part F. References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19*(6), 716-723. doi:10.1109/TAC.1974.1100705

Bell, S. M., Chang, X., Wambaugh, J. F., Allen, D. G., Bartels, M., Brouwer, K. L. R., . . . Kleinstreuer, N. C. (2017). In vitro to in vivo extrapolation for high throughput prioritization and decision making. *Toxicology in Vitro*, in press.

Bell, S. M., Phillips, J., Sedykh, A., Tandon, A., Sprankle, C., Morefield, S. Q., . . . Kleinstreuer, N. C. (2017). An Integrated Chemical Environment to Support 21st-Century Toxicology. *Environ Health Perspect, 125*(5), 054501. doi:10.1289/EHP1759

Burnham, K. P. A. D. R. (2003). *Model selection and multimodel inference: a practical information-theoretic approach.*

Grulke, C. M., Williams, A. J., Thillanadarajah, I., & Richard, A. M. (2019). EPA’s DSSTox database: History of development of a curated chemistry resource supporting computational toxicology research. *Computational Toxicology, 12*, 100096. doi:<https://doi.org/10.1016/j.comtox.2019.100096>

Jones, E., Oliphant, T., & Peterson, P. (2001). SciPy: Open source scientific tools for Python.

Kim, Y. C., Kang, H. E., & Lee, M. G. (2008). Pharmacokinetics of phenytoin and its metabolite, 4'-HPPH, after intravenous and oral administration of phenytoin to diabetic rats induced by alloxan or streptozotocin. *Biopharm Drug Dispos, 29*(1), 51-61. doi:10.1002/bdd.591

Oreskes, N. (1998). Evaluation (not validation) of quantitative models. *Environmental Health Perspectives, 106*(Suppl 6), 1453.

Richard, A. M. (2004). DSSTox Website Launch: Improving Public Access to Databases for Building Structure-Toxicity Prediction Models. *Preclinica, 2*(2), 103-108.

Rohatgi, A. (2018). WebPlotDigitizer (Version 4.1). Retrieved from <https://automeris.io/WebPlotDigitizer>

Sayre, R. R., Wambaugh, J. F., & Grulke, C. M. (2019). *Database of pharmacokinetic time-series data and parameters for XX environmental chemicals*.

Sioutos, N., Coronado, S. d., Haber, M. W., Hartel, F. W., Shaiu, W.-L., & Wright, L. W. (2007). NCI Thesaurus: A semantic model integrating cancer-related clinical and molecular information. *Journal of Biomedical Informatics, 40*(1), 30-43. doi:<https://doi.org/10.1016/j.jbi.2006.02.013>

U.S. Environmental Protection Agency. (1998). *Health Effects Test Guidelines OPPTS 870.7485 MEtabolism and Pharmacokinetics*. Washington, D.C. Retrieved from <http://nepis.epa.gov>

Wambaugh, J. F., Hughes, M. F., Ring, C. L., MacMillan, D. K., Ford, J., Fennell, T. R., . . . Thomas, R. S. (2018). Evaluating in vitro-in vivo extrapolation of toxicokinetics. *Toxicological Sciences, 163*(1), 152-169.

Wambaugh, J. F., Ring, C., & Setzer, W. (2019). invivoPKfit.

Wambaugh, J. F., Wetmore, B. A., Pearce, R., Strope, C., Goldsmith, R., Sluka, J. P., . . . Shah, I. (2015). Toxicokinetic triage for environmental chemicals. *Toxicological Sciences, 147*(1), 55-67.

Williams, A. J., Grulke, C. M., Edwards, J., McEachran, A. D., Mansouri, K., Baker, N. C., . . . Richard, A. M. (2017). The CompTox Chemistry Dashboard: a community data resource for environmental chemistry. *Journal of Cheminformatics, 9*(61). doi:10.1186/s13321-017-0247-6