ToxRefDB v2.0: Building an enhanced resource for predictive applications and beyond

Katie Paul Friedman, PhD

EPA-ORD-CCTE

Team: Sean Watford, Madison Feshuk, and others in CCTE

Updated 2021

Why is legacy data in ToxRefDB 2.0 important to computational toxicology?

- Basis for validation of new approach methods to identify specific adverse outcomes of interest.

- Retrospective benchmark for predictive performance of alternative approaches to predicting quantitative points-of-departure .

- Using ToxRefDB to develop an understanding of the reproducibility and variability in *in vivo* toxicity testing clearly supports development of baseline expectations for new approach methods that promise to assist with rapid prioritization and screening level assessments (Casati et al 2017; Judson et al. 2017; Pham *et al.* 2020).

Accessibility: ToxRefDB v2.0 is a large publicly available resource for computational toxicology research


Reproductive Toxicology 89 (2019) 145–158

Contents lists available at ScienceDirect

Reproductive Toxicology

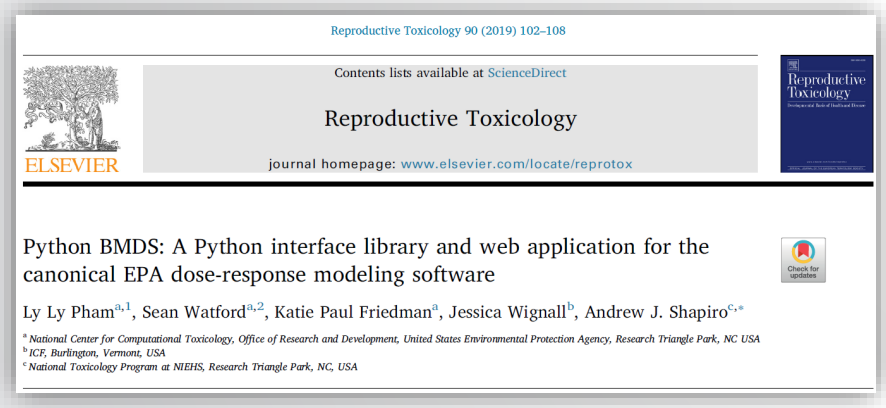
journal homepage: www.elsevier.com/locate/reprotox

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

Sean Watford[a,b], Ly Ly Pham[a,c], Jessica Wignall[d], Robert Shin[d], Matthew T. Martin[a,e], Katie Paul Friedman[b,*]

[a] ORAU, Contractor to U.S. Environmental Protection Agency through the National Student Services Contract, United States
[b] National Center for Computational Toxicology, Office of Research and Development, US Environmental Protection Agency, United States
[c] ORISE Postdoctoral Research Participant, United States
[d] ICF, Burlington, VT, United States
[e] Currently at Drug Safety Research and Development, Global Investigative Toxicology, Pfizer, Groton, CT, United States

https://doi.org/10.1016/j.reprotox.2019.07.012


Reproductive Toxicology 90 (2019) 102–108

Contents lists available at ScienceDirect

Reproductive Toxicology

journal homepage: www.elsevier.com/locate/reprotox

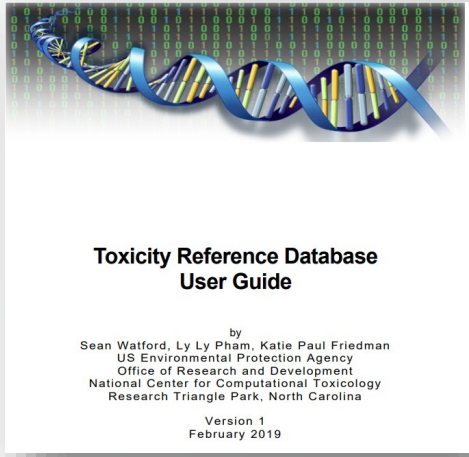
Python BMDS: A Python interface library and web application for the canonical EPA dose-response modeling software

Ly Ly Pham[a,1], Sean Watford[a,2], Katie Paul Friedman[a], Jessica Wignall[b], Andrew J. Shapiro[c,*]

[a] National Center for Computational Toxicology, Office of Research and Development, United States Environmental Protection Agency, Research Triangle Park, NC USA
[b] ICF, Burlington, Vermont, USA
[c] National Toxicology Program at NIEHS, Research Triangle Park, NC, USA

https://doi.org/10.1016/j.reprotox.2019.07.013

The entire database is released as a .sql file that can be mounted with MySQL server, as described in the User Guide.



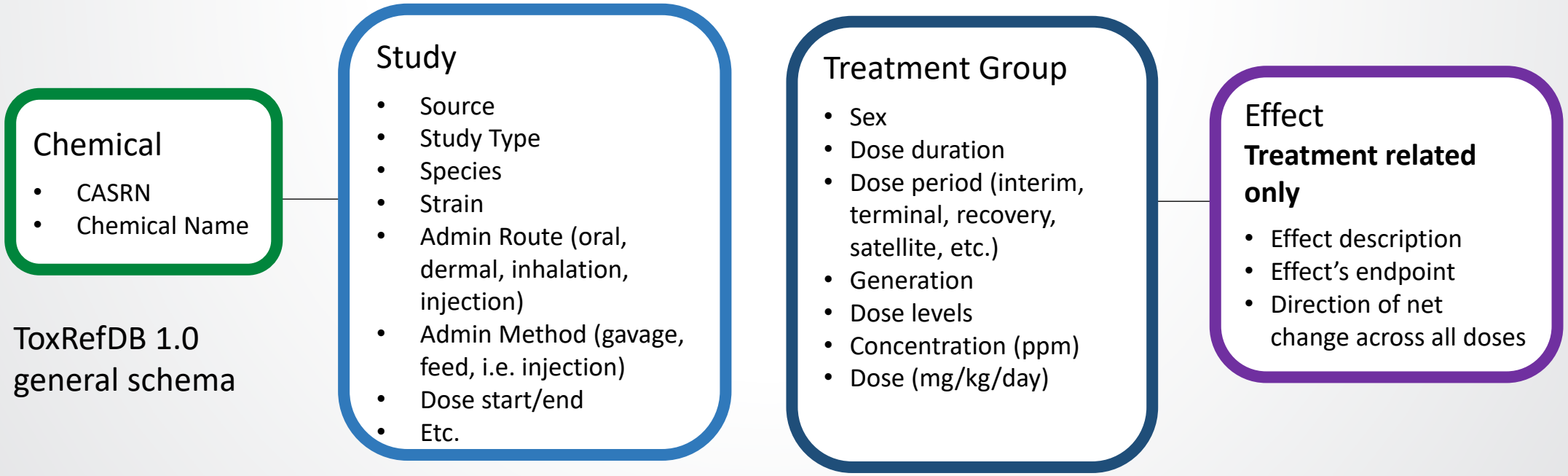
DOI for ToxRefDB v2.0 database:
https://doi.org/10.23645/epacomptox.6062545.v3

Link to download database and associated materials:
https://gaftp.epa.gov/comptox/High_Throughput_Screening_Data/Animal_Tox_Data/current

User guide with code examples for using MySQL to extract data:
https://gaftp.epa.gov/comptox/High_Throughput_Screening_Data/Animal_Tox_Data/current/ToxRefDB_2_0_UserGuide_Final.pdf

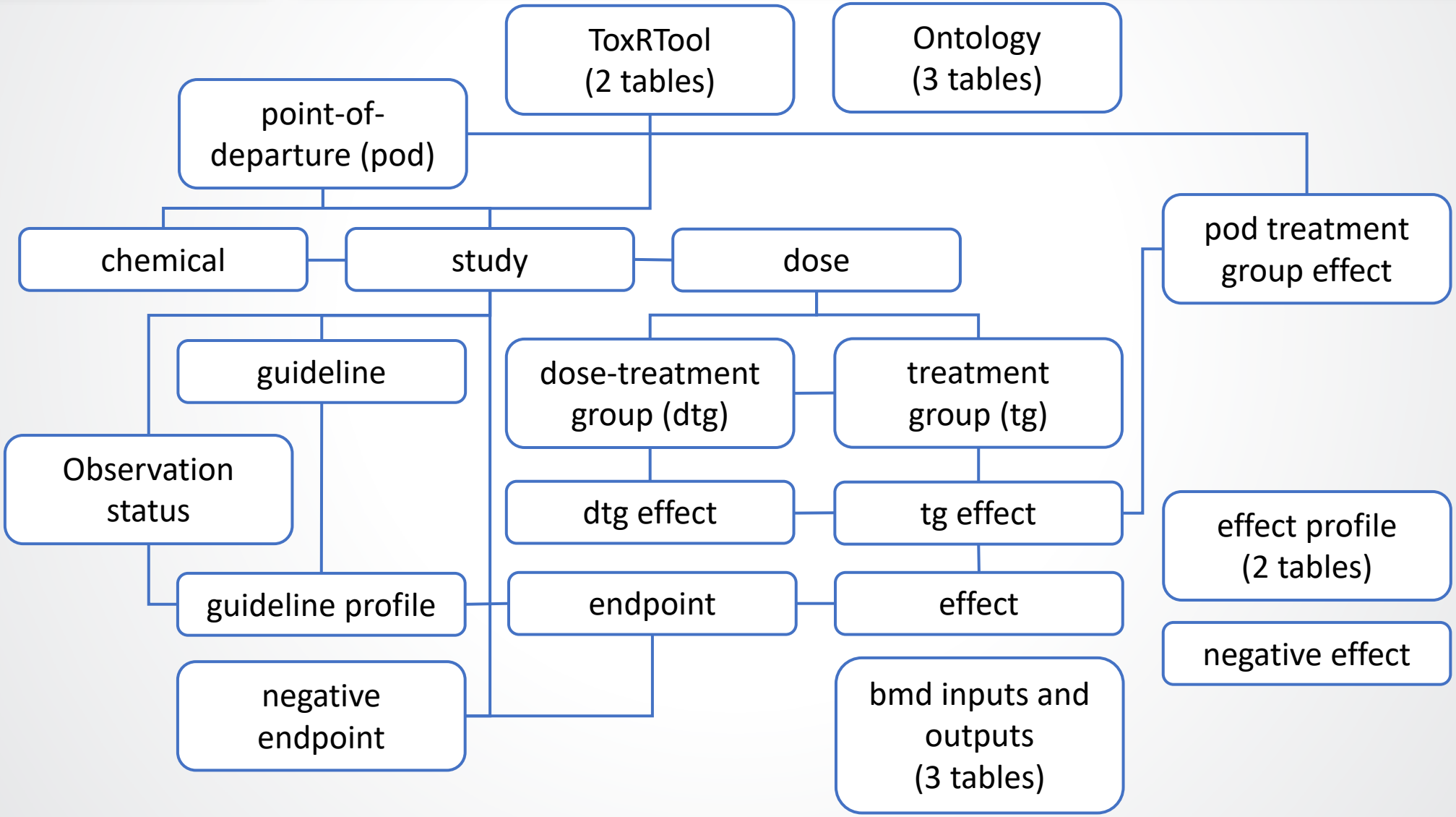
ToxRefDB v1.0: past work

- Priority at its inception – Office of Pesticide Programs (OPP) data evaluation records (summaries of registrant-submitted studies) were going to be a rich data source for comparison to ToxCast Phase I
- Largest public database of *in vivo* toxicology data, with study design and effects
- ToxRefDB 1.0 captured basic study design, treatments, and treat-related effects
  - Positive-only database
  - Mostly qualitative data, only LELs and LOELs



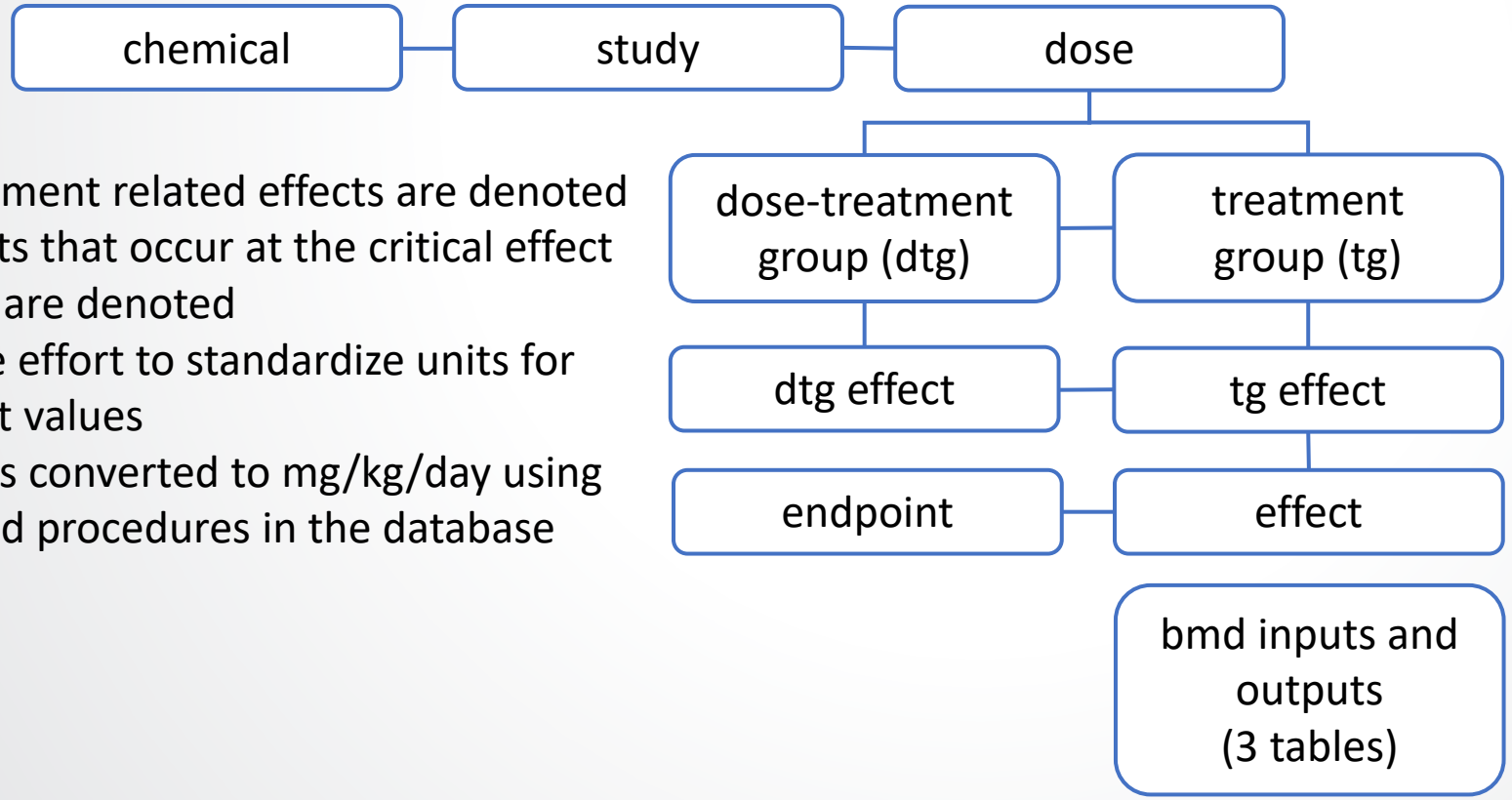
ToxRefDB 1.0 general schema

Generalized ToxRefDB v2.0 schema: added complexity



Quantitative value has been greatly improved

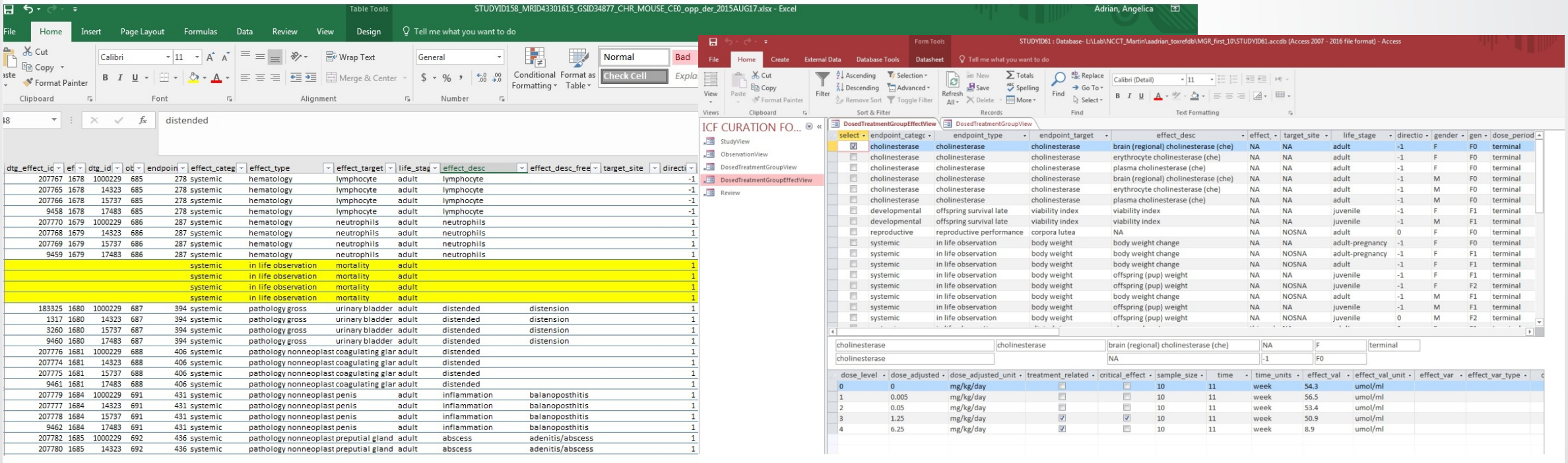
- Treatment related effects are denoted
- Effects that occur at the critical effect level are denoted
- Large effort to standardize units for effect values
- Doses converted to mg/kg/day using stored procedures in the database



- ↑ quantitative value: control and responses at all doses
- ↑ accuracy of mapping of dose and effect to each treatment group (e.g., for studies with multiple generations or male and females)
- Largest implementation of Python-driven BMDS v2.7 to provide BMDL, BMD, and BMDU values from winning models whenever practicable

Quantitative extraction via Access form entry

- Dose-treatment group-effect quantitative data
- Switched from Excel sheet entry to Access form entry
  - Only treatment-related effects entered into ToxRefDB 1.0, so no control groups
  - Control groups manually entered, which lead to human errors
  - Control groups automatically generated in Access files



Access forms decrease error rate; extraction and import processes have additional QA steps

At data entry (at ICF using our Access files)	At data import (at EPA-CCTE)
Primary reviewer • Extracts per instructions Secondary reviewer • Confirms each piece of information from 1st extraction • Reference comment log as needed Senior toxicology reviewer • Review extractions and comments from primary and secondary reviewers	Automated checks for the following errors: • Mismatched dose levels with concentration or dose-adjusted • Duplicate concentration values or duplicate controls • No concentration and no dose adjusted value (for an effect) • Critical effect level is at a dose below where treatment-related effects were observed or critical effect at control • NULL concentrations/doses

Automated BMDS pipeline: additional quantitative value for future analyses

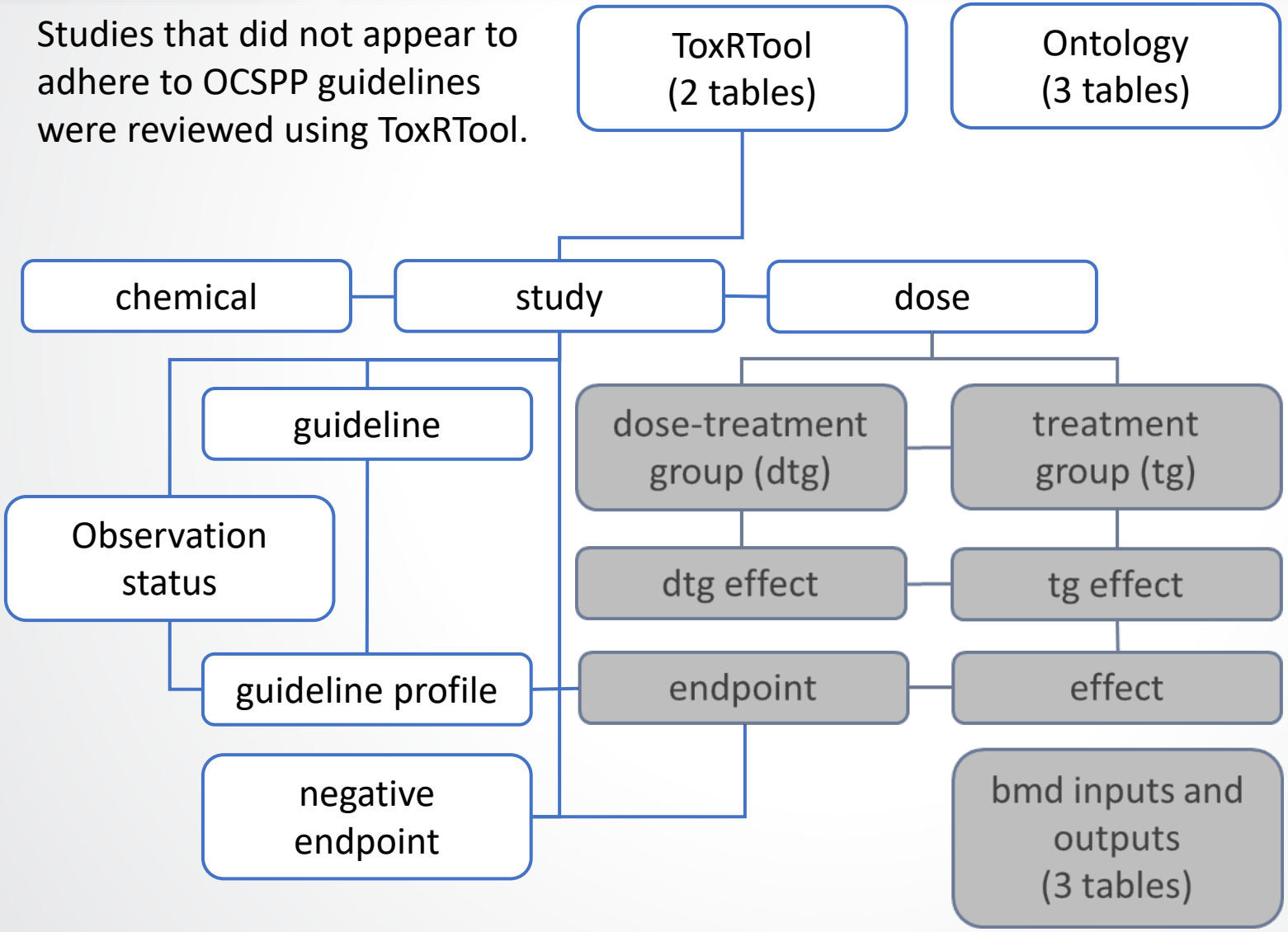
- Calculate BMD and BMDL (and BMDU) for all treatment-related effects that pass minimum data requirements for BMDS
- Batch BMDS v2.7 with python package bmds (https://github.com/shapiromatron/bmds)
- Currently, there are 92,646 datasets in ToxRefDB with at least 3 doses from "acceptable" studies, but only ~30% of them are BMDS-amenable at this time.
- 6,231 out of 27,757 BMDS-amenable datasets had at least one recommended model.
- Application note that describes the use of a Python-driven BMDS pipeline is in preparation (Pham *et al.*).

Data type	Number of datasets available for BMDS	Benchmark responses used*
Cancer	1,340	5% and 10%
Non-cancer dichotomous	17,149	5% and 10%
Continuous body/organ weight	9,268	10% relative deviation
Continuous non-body/organ weight		1 standard deviation

* *Based on recommendations in the BMDS guidance (2012)*

Controlled effect and endpoint vocabulary has enabled a number of improvements and interoperability.

- Studies that did not appear to adhere to OCSPP guidelines were reviewed using ToxRTool.



- Currently we have the ToxRefDB ontology mapped to the United Medical Language System.

- The ToxRefDB vocabulary has been updated.
- Guideline, guideline profile, observation status all enable automated generation of true negative endpoints and effects.

negative effect

OCSPP guidelines and NTP study guidance were used to generate “guideline profiles” in the database

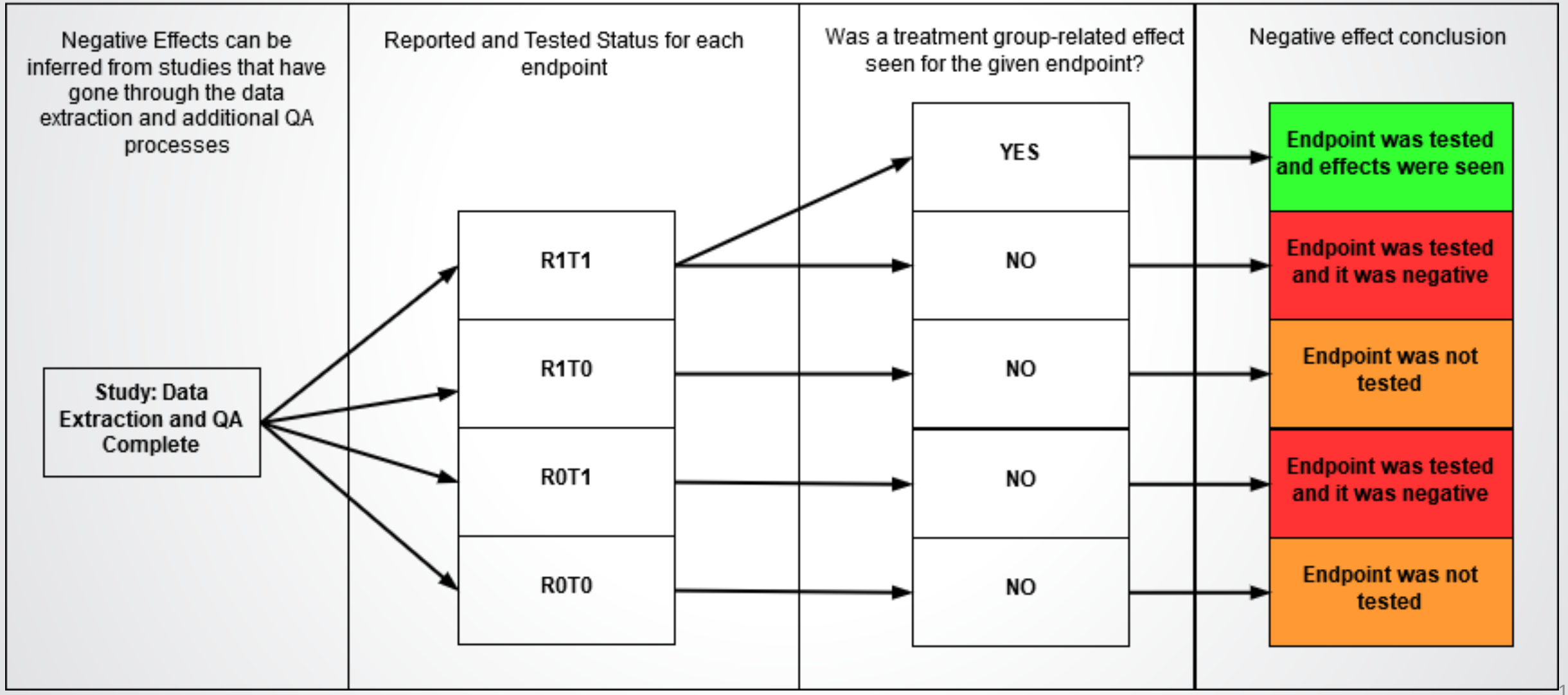
- Endpoint testing requirements as indicated by OCSPP 870 series guidelines or NTP specifications
  - Other subsources cannot be uniformly mapped
- Allows for default assumptions about testing and reporting for inference of true negatives
- Required, triggered, recommended, not required

Observation status enables automated distinction of true negatives

- Allows for distinction between missing (not tested) and negative (tested with no effect seen) endpoints
- Assumes that an endpoint was tested if the guideline the study adheres to requires that endpoint be tested
- Defaults in access files check both tested and reported if the guideline requires it

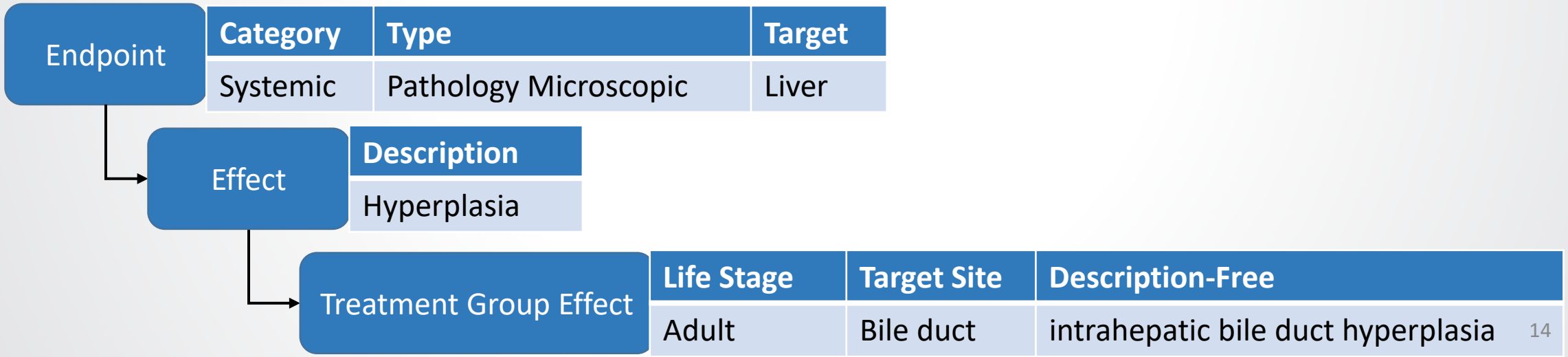
Tested status	Reported status	Case in the database
Tested	Reported	The endpoint was SPECIFICALLY written in the text of the study source indicating that data was collected (default if required by the guideline for that study type and no deviation reported)
Not tested	Reported	The endpoint was SPECIFICALLY 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 the information 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

General workflow to infer negative endpoints/ effects



ToxRefDB vocabulary update

- Fixed problems with vague terminology that created duplicate endpoints
  - Terminology standardized to match OCSPP Testing Guidelines
- Simplified the pathological concepts
  - Pathology gross and microscopic instead of carcinogenic neoplastic and nonneoplastic
- Now have ~400 endpoints from originally ~500



Clinical Data Interchange Standards Consortium (CDISC) wasn't quite expansive enough for us

- Standards developing organization to streamline medical research
- Partners with NCI to develop and support controlled terminology
  - SDTM: Study Data Tabulation Model - Provides recommended standards for human and nonclinical data submitted to FDA
  - **SEND: Standard for Exchange of Non-Clinical Data - Provides standardized terminology for non-clinical (animal) data**
  - CDASH: Clinical Data Acquisition Standards Harmonization – Develops clinical research study content standards
  - ADaM, Protocol, LAB, others
- FDA, Pharmaceuticals and Medical Devices Agency (Japan), European Medicines Agency (EMA), and eTOX all use CDISC
- Common ontology will allow for interagency exchange of data

United Medical Language System (UMLS) cross-references 150 vocabularies

- National Library of Medicine (NLM)
- National Cancer Institute Thesaurus (NCIt) includes ~90 controlled vocabularies and is a subset of UMLS
- CDISC terminology is submitted to NCIt
- Concepts (CUI or NCI Code) < atoms (instance of the concept in a particular system)
- ~3M concepts with ~7M atoms

Search | Tree | Recent Searches

Term | CUI | Code

liver Go

Release: 2017AB

Search Type: Word

Source: All Sources, AIR, ALT, AOD, AOT

Search Results (4577)

[: 1 - 25 :]

C0023884 Liver
C0023899 Liver Extract
C0721399 Liver brand of Vitamin B 12
C1278929 Entire liver
C2346688 Liver Flavor
C0745746 LIVER DISEASE SHOCK LIVER
C0013504 Echinococcosis, Hepatic
C0015648 Fasciola hepatica
C0015695 Fatty Liver
C0017796 Glutaminase
C0017911 Glycogen
C0019144 Hepatectomy
C0019158 Hepatitis
C0019209 Hepatomegaly
C0019213 Suture of liver
C0022801 Hepatic macrophage
C0023885 Liver Abscess
C0023889 Liver Circulation
C0023890 Liver Cirrhosis
C0023895 Liver diseases
C0023896 Alcoholic Liver Diseases
C0023901 Liver Function Tests
C0023902 Liver Glycogen
C0023903 Liver neoplasms

Basic View | Report View | Raw View

Concept: [C0023884] Liver

Semantic Types

Body Part, Organ, or Organ Component [T023]

Definitions

CSP | large gland of a dark-red color situated in the upper part of the abdomen on the right side; comprised of thousands of minute lobules, the functional units of the liver; functions include the storage and filtration of blood, secretion of bile, excretion of bilirubin and other substances formed elsewhere, and numerous metabolic functions, including the conversion of sugars into glycogen, which it stores.

FMA | Lobular organ which has as its parts lobules connected to the biliary tree. Examples: There is only one liver.

MSH | A large lobed glandular organ in the abdomen of vertebrates that is responsible for detoxification, metabolism, synthesis and storage of various substances.

MSHNOR | Et stort, lappet abdominalorgan hos virveldyr og mennesker som sørger for metabolisme, syntese og lagring av forskjellige substanser, samt nøytralisering av toksiner.

NCI | A large organ located in the upper abdomen. The liver cleanses the blood and aids in digestion by secreting bile.

NCI | A triangular-shaped organ located under the diaphragm in the right hypochondrium. It is the largest internal organ of the body, weighting up to 2 kg. Metabolism and bile secretion are its main functions. It is composed of cells which have the ability to regenerate.

NCI | An abdominal organ that has variable lobation which are composed mainly of hepatic lobules.

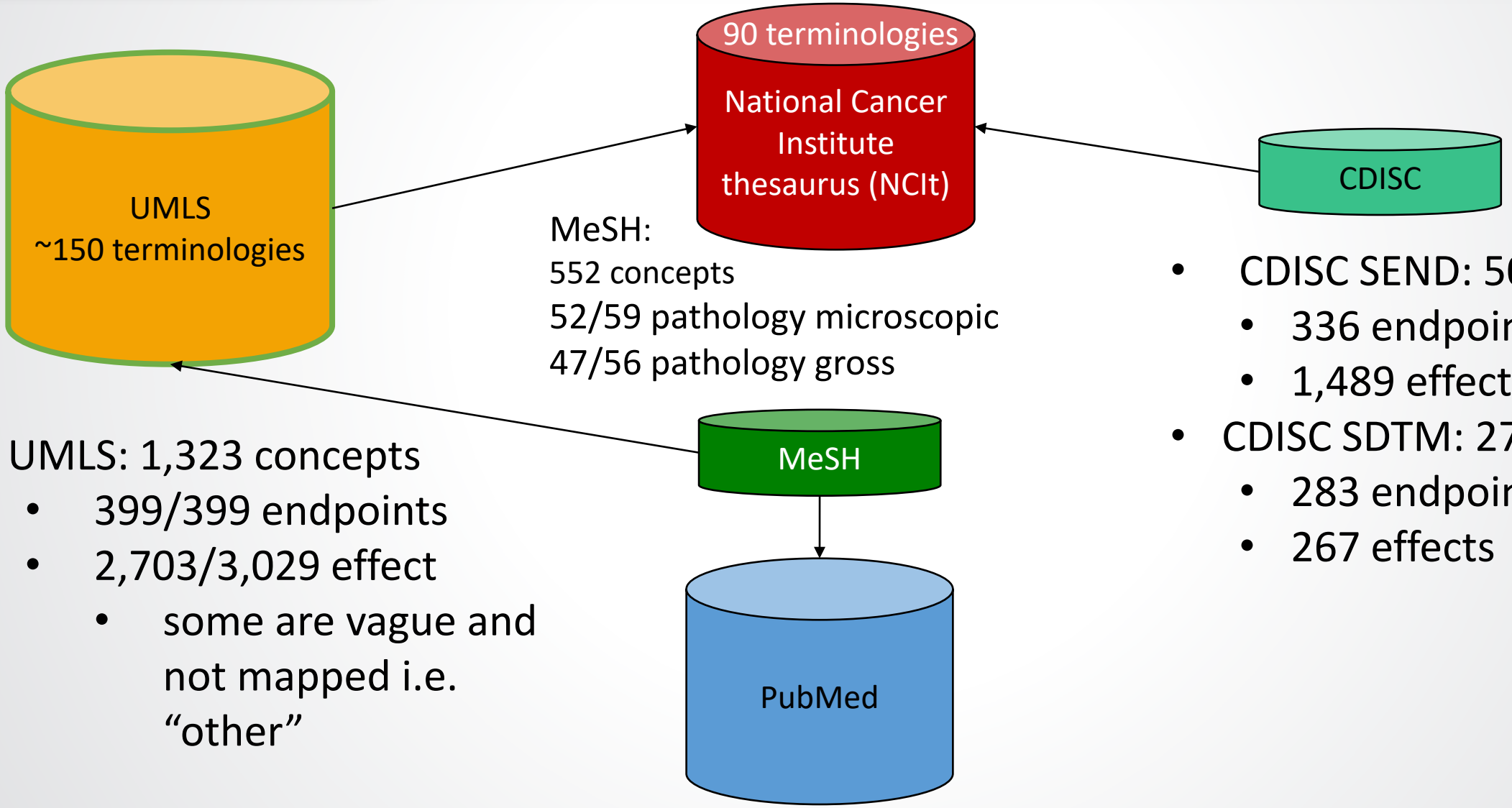
UWDA | Lobular organ the parenchyma of which consists of lobules which communicate with the biliary tree. Examples: There is only one liver.

Atoms (73) string [AUI / RSAB / TTY / Code]

liver [A0482918/AOD/DE/0000002569]
LIVER [A0428018/CCPSS/PT/0050526]
liver [A18573838/CHV/PT/0000007494]
liver structure [A18592464/CHV/SY/0000007494]
livers [A18592465/CHV/SY/0000007494]
liver [A0482919/CSP/PT/1754-0095]
Hepar [A15400356/FMA/SY/7197]
Liver [A15467341/FMA/PT/7197]
Liver [A8311874/HL7V2.5/PT/LIVER]
Structure of liver [A16037325/ICF/PT/s560]
Structure of liver [A16038998/ICF-CY/PT/s560]
Liver [A0080669/LCH/PT/U002724]
Liver [A23865285/LCH_NW/PT/sh85077748]
Liver [A18168047/LNC/LPN/LP29289-3]
Liver [A18329241/LNC/LPN/LP7400-7]

Endpoint Category	Endpoint Type	Target
systemic	pathology microscopic	liver

Summary of mapping ToxRefDB to UMLS

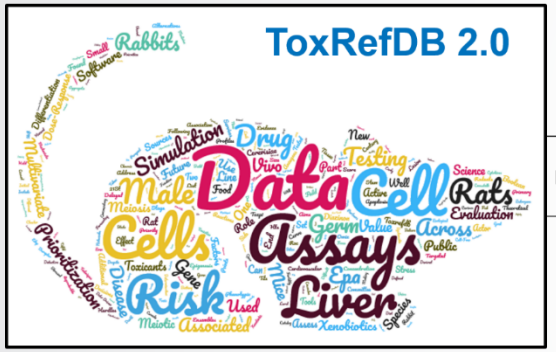


MeSH:
552 concepts
52/59 pathology microscopic
47/56 pathology gross

- UMLS: 1,323 concepts
  - 399/399 endpoints
  - 2,703/3,029 effect
    - some are vague and not mapped i.e. “other”

- CDISC SEND: 500 concepts
  - 336 endpoints
  - 1,489 effects
- CDISC SDTM: 279 concepts
  - 283 endpoints
  - 267 effects

Implications of UMLS mapping for data integration with public biomedical sources



Endpoints/Effects

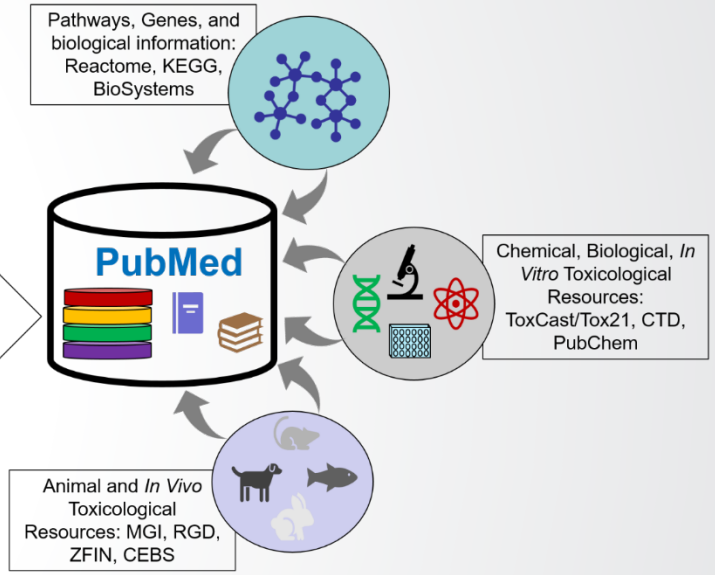
Category	Type	Target	Description	UMLS Concept Name
systemic	pathology microscopic	mammary gland	adenocarcinoma	Adenocarcinoma
systemic	pathology microscopic	liver	adenocarcinoma	Adenocarcinoma
systemic	pathology microscopic	kidney	adenocarcinoma	Adenocarcinoma
systemic	pathology microscopic	liver	adenoma/carcinoma combined	Adenoma
systemic	pathology microscopic	mammary gland	adenoma/carcinoma combined	Adenoma
systemic	pathology microscopic	liver	adenoma	Adenoma
systemic	pathology microscopic	kidney	adenoma/carcinoma combined	Adenoma

Concepts

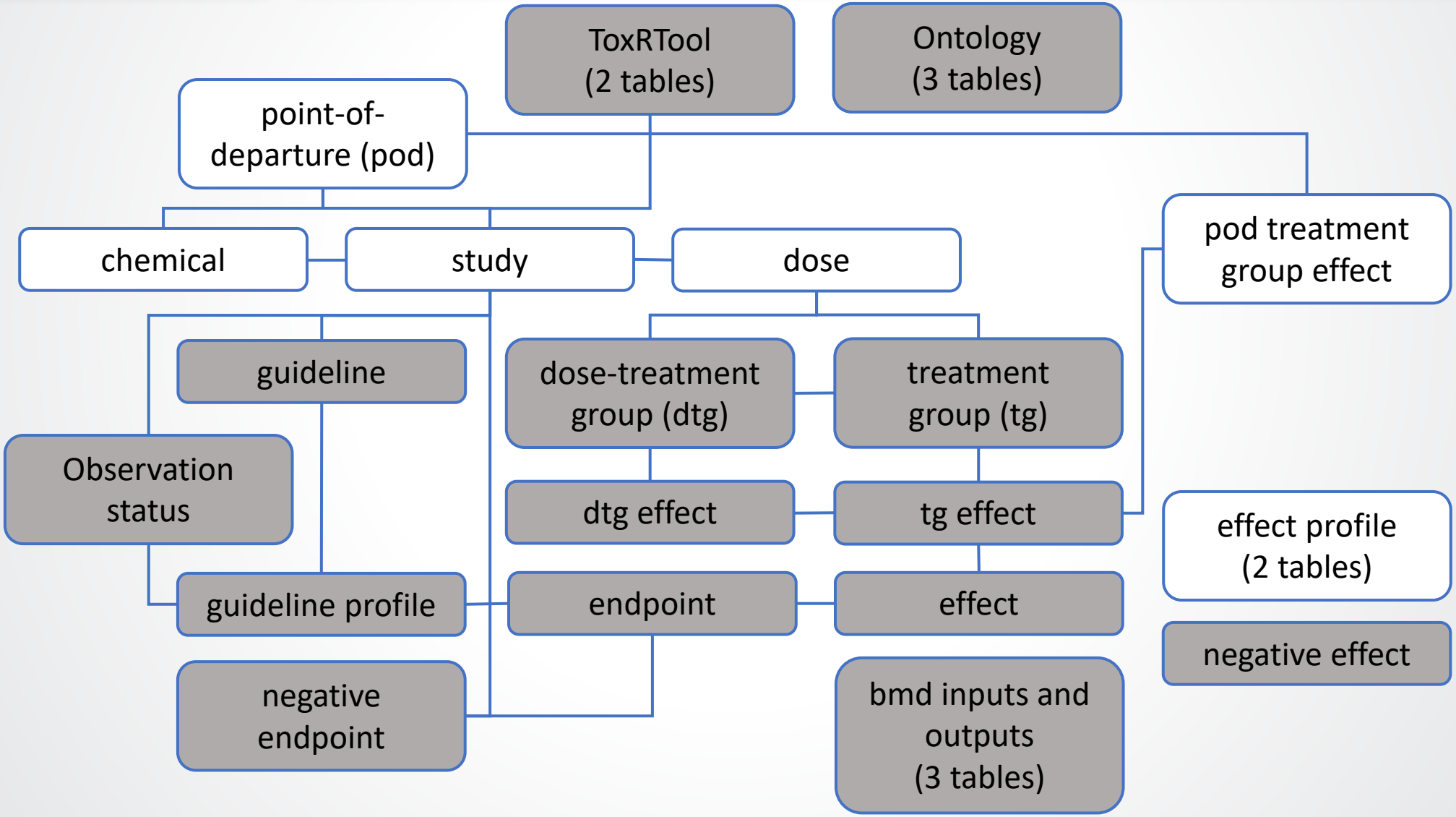
UMLS

NCI Thesaurus | CDISC-SEND | Other Terminologies

MeSH



Calculation of points-of-departure by study and by grouping of effects



Point-of-departure estimates can be computed in different ways based on use case

- The way it is set up, we can add more “effect profiles” to group effects for computation of PODs
- Study level POD: use the treatment-related effects and critical effect to get the NOAEL and LOAEL (and NEL and LEL) BY study
- Chemical level POD: report multiple NEL/LEL/NOAEL/LOAEL sets if there are effects in multiple domains
  - Currently, we used the endpoint_category level (i.e., cholinesterase, developmental, reproductive), except when endpoint_category=systemic, we used endpoint_target (e.g., liver, clinical signs, in-life observations)
  - Chemical level PODs go to ToxValDB