SoT 2023 Annual Meeting Abstract #5109/P210

Data Profiling and Quality Control Pre-Screening of Toxicology Data in ToxValDB

Mitchell C. Tague[2], Anthony Brito[2], Richard Judson[1], Evelyn Rowan[2], Taylor Wall[1], Risa R. Sayre[1]

1. Center for Computational Toxicology and Exposure (CCTE), U.S. Environmental Protection Agency, 109 T.W. Alexander Dr., Research Triangle Park, NC 27709, USA.
2. Oak Ridge Associated Universities, Oak Ridge Institute for Science and Education, Oak Ridge, TN, USA. Contracted for US EPA ORD/CCTE

Mitchell Tague | tague.mitchell@epa.gov | Phone: 919-541-1063 | ORCID: 0000-0003-4046-235X

Background

ToxValDB is a database of quantitative data including but not limited the results of *in vivo* toxicology studies, risk screening levels, and reference doses. To ensure EPA researchers and outside partners have clean data, our team is developing a continuous quality control workflow for ToxValDB data. Due to the inherent difficulties of data quality assurance when aggregating from several sources, data profiling was a needed first step.

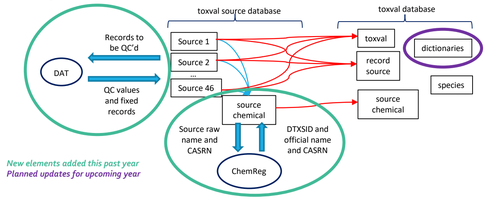


Figure 1: Structure of ToxValDB construction. DAT: Data Accuracy Tool, an EPA application for easy record QC; ChemReg: an EPA app for chemical curation

Workflow

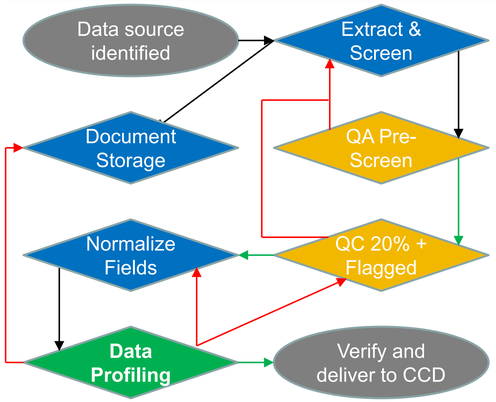


Figure 2: Workflow to add data into and QC data from ToxValDB. The green diamond represents this poster's content: profiling the data within the finalized ToxValDB to ensure data quality. Yellow diamonds indicate QC processes performed by a domain specialist wherein records are checked against the original source. Blue diamonds are steps within the workflow performed by a data scientist. Green and red arrows represent QC passes or failures. The CCD stands for the "CompTox Chemical Dashboard" (see References).

Data Profiling

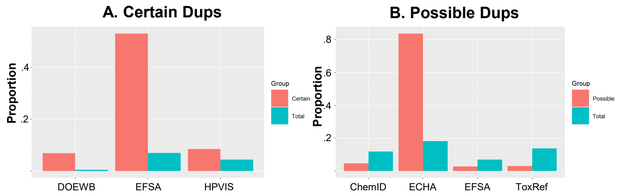
A. Duplicates

Motivation: Presence of duplicate entries detected, potentially from entry errors, duplication in sources, and same studies provided by separate sources.

Results:
- 25,891 (6.29%) "certain" dups.
- 72,792 (17.69%) "possible" dups.
- Heavily source dependent – EFSA dominates "certain" dups, ECHA IUCLID for "possible" dups.

Process: Identify categories of duplicates for flagging:
1. "Certain" duplicates – records which match in all fields outside of record identifier fields.
2. "Possible" duplicates – records which match in the most study-relevant fields.

Manual review of QC will allow per-source updating of fields which define "possible" dups.



Figures 3a and 3b: Sources overrepresented in proportions of duplicate records. Each graph compares the proportion of the source comprised of "certain" or "possible" duplicates with the overall proportion of dups in the total dataset. 3a (left) looks at *certain* duplicates, for sources overrepresented in the *certain* duplicates. 3b (right) plots sources containing at least 2% of the total set of *possible* duplicates.

B. Normalization Errors

Motivation: Many fields in ToxValDB are aggregated into normalized values for cross-source analyses. Investigate for further normalization needs, and ensure prior normalization is valid.

Process: Check values with Levenshtein distance = 1 within a given field to identify further normalization needs. Check cases where differing original values normalize to a single value, to ensure validity.

Results: There are no cases where entries with similar original values are normalized into different categories. All original values that were normalized together appear logical. Similar values among the normalized forms of some fields (e.g. "adult" vs "Adult", "Sprague Dawley Rat" vs "Sprague-Dawley Rat) suggests more normalization is needed for population, lifestage, and study duration class.

C. Numeric Profiling

Motivation: High spread in numeric values identified within certain field combinations. Potential for numeric errors both within sources and due to extraction. Numeric values highly dependent on other fields, yielding difficulty in analysis.

Process: Group records with matching toxval type and subtype, chemical, units, and species. Profile for outliers, groups with large spread, and high repeat groups. Repeat process for type and units only, on data unable to be analyzed by more complete field combination.

Results
- 108890 (35.3%) of records analyzed by full combination.
- 9394 (3.04%) outliers flagged. 7337 outliers identified in full combination.
- 3181 large spread group records flagged.
- 13305 replicate group records – limit tests, likely no issues.

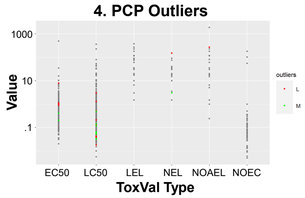
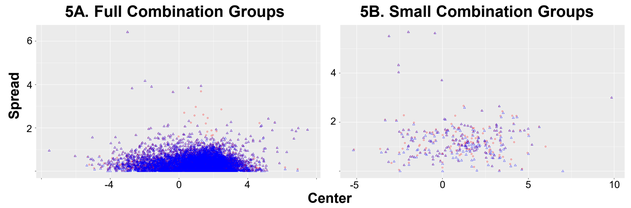


Figure 4 (left): Pentachlorophenol (PCP) values for select toxval types, with large ("L") and medium ("M") outliers selected.
Figures 5a (below, left) and 5b (below): Scatterplot of groups from profiling, showcasing center (x) and spread (y) of log-transformed data. 5a represents full group overlaps, 5b represents small group overlaps. Blue triangles indicate pre-outlier removal, red circles represent post-outlier removal.



D. Other Issues

- 443 prior QC fails were identified to have multiple units.
- ToxVal Types listed as "AEGL" contained extra information useful in other fields. 3,333 records flagged for review.
- 436 records flagged for being low frequency type (e.g. LD70), 470 records flagged for having low frequency units (e.g. uL/kg-day). 218 separate records could not be analyzed due to lack of comparison data.

Results

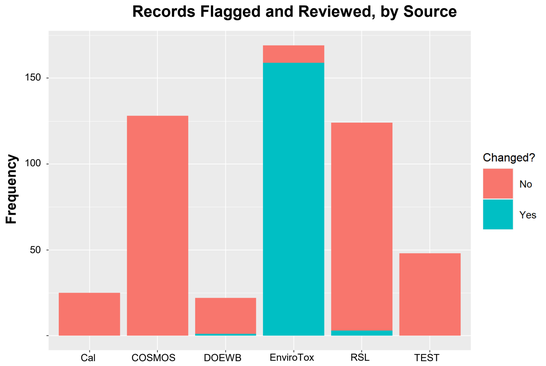


Figure 6: Results (by source) of manual quality check of subset of flagged documents, indicating whether flagged records were manually confirmed to have errors compared to source. Five percent per-source of records flagged as duplicate were manually reviewed, and 5% per-source of records flagged via numeric profiling that were also about a chemical on the TSCA Active Inventory list were reviewed. Certain sources with known issues excluded from review. Only sources with at least 20 records reviewed are included in the figure.

Discussion

Looking at which source a record comes from proved to be a better indicator of the need for manual curation than any data profiling flag.
- Certain duplicates in EFSA and DOEWB sources already addressed.
- Most possible duplicates reviewed (86.5%) were not true duplicates. Differentiating field depends on source.
- 51/71 (71.8%) of analyzed outliers from full combination required updates – higher than any other flag type.
- Only 5/20 (25%) of analyzed outliers from small combination required updates. Analysis by full overlap alone recommended.
- Continued data loads to ToxValDB require continuous profiling.

References

- ECHA, IUCLID 6, https://iuclid6.echa.europa.eu/get-iuclid-data
- Kovarich, S. et al. (2022). OpenFoodTox: EFSA's chemical hazards database (Version 5) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.5076033
- U.S. EPA, CompTox Chemical Dashboard https://comptox.epa.gov/dashboard/